Linköpings universitet | Institutionen för datavetenskap
729G40 Kandidatuppsats, 12 hp | Kognitionsvetenskap
Vårterminen 2019 | LIU-IDA/KOGVET-G--19/017--SE

Fördomsfulla associationer i en svensk vektorbaserad semantisk modell

Michael Jonasson

Handledare: Fredrik Stjernberg
Examinator: Leelo Keevallik

LINKÖPINGS UNIVERSITET

Abstract/Sammanfattning

Word embeddings are a powerful technique where word meaning can be represented by vectors containing actual numbers. The vectors allow geometric operations that capture semantically important relationships between the words. In this study WEAT is applied in order to examine whether statistical properties of words pertaining to bias can be found in a Swedish word embedding trained on a corpus from a Swedish newspaper. The results shows that the word embedding can represent several of the IAT documented biases that where tested. A second method, WEFAT, is applied to the word embedding in order to explore the embeddings ability to represent actual statistical properties, which is also done successfully. The results from this study lends support to the validity of both methods aswell as illuminating the issue of problematic relationships between words in word embeddings.

Semantiska vektormodeller är en kraftfull teknik där ords mening kan representeras av vektorer vilka består av siffror. Vektorerna tillåter geometriska operationer vilka fångar semantiskt viktiga förhållanden mellan orden de representerar. I denna studie implementeras och appliceras WEAT-metoden för att undersöka om statistiska förhållanden mellan ord som kan uppfattas som fördomsfulla existerar i en svensk semantisk vektormodell av en svensk nyhetstidning. Resultatet pekar på att ordförhållanden i vektormodellen har förmågan att återspegla flera av de sedan tidigare IAT-dokumenterade fördomar som undersöktes. I studien implementeras och appliceras också WEFAT-metoden för att undersöka vektormodellens förmåga att representera två faktiska statistiska samband i verkligheten, vilket görs framgångsrikt i båda undersökningarna. Resultaten av studien som helhet ger stöd till metoderna som används och belyser samtidigt problematik med att använda semantiska vektormodeller i språkteknologiska applikationer.

Innehållsförteckning

Inledning **5**
Syfte och frågeställning 7
Avgränsningar 7
Tidigare forskning: **8**
Teori **9**
Implicit kognition 9
Omedveten partiskhet och fördomsfullt beteende 10
IAT 10
IAT procedur *10*
Data och resultat från IAT *11*
Kritik mot IAT *11*
Distribuerade semantiska modeller 13
Användningsområden *13*
Begränsningar med distribuerade semantiska modeller *14*
Word2vec *15*
Människan och distribuerad semantik *15*
Metod **17**
WEAT 17
Cosinusmått *17*
Associationsstyrka mellan flera ord *17*
WEAT *18*
Effektstorleken av WEAT *18*
Permutationstest *19*
WEFAT 19
Material **21**
Vektormodellen 21
Datainsamling 21
IAT-resultat *21*
Data från statistiska undersökningar *22*
Resultat **24**
Diskussion **27**
Framtida forskning 28
Slutsats 29
Källor: **30**
Bilaga 1 ..... **34**
Bilaga 2 - Yrkeskategorier **40**

Inledning

Att förstå språk handlar delvis om att förstå de regler som ligger bakom kombinationen av ord, och delvis av att förstå ordens innebörd (Bermudez, 2012). Att använda datorer för att skapa meningar som följer de grammatiska regler som finns är något som har gjorts sedan 1960, då med program som exempelvis ELIZA (Weizenbaum, 1966) och SHRLDU (Winograd, 1971) som framgångsrikt kunde imitera naturligt språk utan att förstå semantiken bakom. Men med stöd från exempelvis lingvistiska teorier som *distributionshypotesen* har en ny typ av maskininlärningsteknik växt fram som har lett till att datorer kan representera ords mening enbart genom att titta på i vilken kontext orden används.

Vektorbaserade semantiska modeller (härefter *vektormodeller*) är en sådan typ av maskininlärningsteknik, där varje ord representeras med vektorer. Dessa vektorer består av siffror som korrelerar mot olika punkter i en vektorrymd (eng. *vector space*). De geometriska förhållandet mellan dessa vektorer fångar meningsfulla semantiska förhållanden mellan de motsvarande orden. Vektorer som existerar nära varandra i vektorrymden överensstämmer med ord som är semantiskt lika. I en exempelmodell så är vektorerna för orden "Playstation" och "Amiga" de som ligger närmast "Xbox", och närmast ordet "Redish" ligger vektorerna för "Bluish" och "Greenish" (Collobert et al., 2011). Då ordvektorena består av siffror går det att applicera matematiska beräkningar på dem; skillnaden mellan "London" och "England", som erhålls genom att subtrahera dessa vektorer, motsvaras av vektorskillnaden mellan "Paris" och "Frankrike". Denna typ av mönster tillåter vektormodeller att fånga analogier så som "London är för England vad Paris är för Frankrike" (Garg, Schiebinger, Jurafsky, & Zou 2018). Att kunna extrahera och representera semantisk innebörd hos ord har ett brett användningsområde och kan användas för allt från att förbättra sökmotorer (Nalisnick, Mitra, Craswell & Caruana 2017) till att analysera CVs (Tosik, Lygteskov-Hansen, Goossen & Rotaru, 2015) och sentimentanalys (Irsoy & Cardie, 2014).

Vektormodellens förmåga att fånga semantiska förhållanden har även en baksida, då dessa semantiska egenskaper kan beskrivas som fördomsfulla och riskerar introducera *bias*[1] i en

[1] Ordet bias saknar en svensk direktöversättning; Psykologiguiden (n, d) definierar bias som "Partiskhet, förutfattad mening, *fördom*, vinkling, så kallad kognitiv bias"

mängd olika applikationer i världen. En av de första studierna som undersöker fördomsfulla associationer i vektormodeller (Bolukbasi, Chang, Zou, Saligrama & Kalai 2016) belyser detta med exemplet att "Man is to computer programmer as woman is to homemaker". Andra man/kvinna förhållanden som identifieras i studien är bland annat *cosmetics-pharmaceuticals, diva-superstar* och *nurse-surgeon*. Vektormodellen som används i studien beskrivs som en populär, allmänt tillgänglig modell tränad på artiklar från *Google news*.

Det implicita associationstestet (eng. *Implicit Association Test; IAT*) har sedan dess skapelse 1998 använts för att undersöka de implicita associationer som tros ligga bakom fördomsfullt beteende (Greenwald, McGhee & Schwartz 1998; Greenwald, Nosek & Banaji 2003). Testet visar att det är lättare för personer att para ihop koncept som de finner lika på en omedveten nivå, än koncept de finner olika. I en amerikansk studie från 2017 lyckades forskare återskapa tidigare dokumenterade IAT-resultat i en vektormodell (Caliskan, Bryson & Narayanan, 2017a). Modellen som används i studien beskrivs återigen som en populär, allmänt tillgänglig modell och är tränad på en korpus med nyhetsartiklar likt ovan. I studien lyckades forskarna replikera samtliga 10 bias som undersöktes, vilka bland annat innefattade moraliskt neutrala bias som att människor anser blommor vara mer behagliga än insekter, till den mer stereotypa[2] biasen att män förknippas mer med vetenskap och kvinnor med konst. Vidare utvecklade forskarna en metod som lät dem återskapa statistiska samband i vektormodellen. I vektormodellen kunde man bland annat se hur olika namns associationsstyrka[3] till män respektive kvinnor återspeglade det faktiska statistiska förhållandet i fördelningen av namnen hos män respektive kvinnor.

De metoder som utvecklats för att undersöka fördomsfulla relationer mellan ord i vektormodeller kan vara viktiga verktyg inte bara för att undersöka potentiella problemområden i en vektormodell, utan även möjliggöra en ny typ av undersökningar där värdefull information om världen kan extraheras enbart genom att titta på hur språk används i stora textsamlingar. Vidare kan kanske dessa metoder användas för att ge ny sorts inblick i fördomar och hur de uppstår.

---

[2] *Stereotyp* inom socialpsykologin: förenklad, ofta allmänt omfattad föreställning om utmärkande egenskaper hos alla som tillhör en viss grupp, till exempel. nation, ras, religion eller kön, också en där man själv ingår.

[3] *Associationsstyrka* innebär styrkan på associationen mellan två eller fler ord baserat på cosinusmått.

Syfte och frågeställning

I studien av Caliskan et al. (2017a) lyfts hypotesen att alla mänskliga implicita bias finns reflekterade i statiska förhållanden hos ord. Det huvudsakliga syftet med denna studie är att fortsatt undersöka denna hypotes. För att göra detta kommer denna rapport att använda de metoder som utvecklats i ovan nämnda studie för att undersöka andra problematiska associationer mellan ord på ett annat språk än i orginalstudien. Förhoppningen är att resultatet från denna studie inte bara ger validitet till både hypotesen och metoderna, utan även belyser eventuella problematiska förhållanden mellan ord i den specifika vektormodell som undersöks. De metoder (WEAT och WEFAT) som används i denna studie kommer implementeras med Python av författaren till denna rapport, vilket innebär att en del av arbetet avser utvecklingen av programmet. Det program som skapats under rapportens arbete bifogas som bilaga.

I denna rapport undersöks två metoder för att arbeta med vektormodeller. Dessa metoder bidrar till ett nytt sätt att arbeta med vektormodeller. Denna rapport ämnar således undersöka frågorna:

1. Går det att identifiera förhållanden mellan ord som liknar fördomar i en svensk vektormodell?
2. I vilken utsträckning kan en svensk vektormodell återspegla faktiska statistiska förhållanden i Sverige?

Avgränsningar

Denna studie avser att undersöka en svensk vektormodell med data insamlad från svenska statistiska- och IAT-undersökningar. För att kunna undersöka om fördomsfulla associationer existerar i den svenska vektormodellen så kommer enbart de associationer som finns dokumenterade i svenska IAT-studier att undersökas. En vektormodell är ett sätt att representera ords semantiska innebörd; denna rapport avser inte att tillföra något till den generella diskussionen gällande en dators förmåga att representera semantik. Den semantik som vektormodellen i den här studien kan representera härstammar från en korpus från en nyhetstidning och kan inte ses som någon allmängiltig semantik för Sverige. Avslutningsvis så har inte författaren till denna rapport skapat den semantiska vektormodell som används i studien, och denna rapport avser därför inte diskutera vektormodeller på någon djupare teknisk nivå.

Tidigare forskning:

Studien av Caliskan et al. (2017a) som nämndes i inledningen är av särskilt intresse. I studien undersöks ett stort spektrum av kända bias genom att göra beräkningar med en vektormodell. De fördomar eller bias som undersökts har tidigare uppmätts av IAT, ett omdebatterat verktyg för att mäta de implicita associationer som enligt många tros lika bakom fördomar. För att kunna genomföra sin studie så utvecklar författarna metoden WEAT (*word embedding association test*), vilket beskrivs som ett statistisk test som är analogt med IAT. Testet undersöker hur starkt associerade två mängder med ord är med två andra mängder ord genom att titta på deras avstånd i förhållande till varandra. Totalt använder sig studien av IAT-resultat från 10 olika studier och hittar signifikanta resultat i 9 av dem med hjälp av den utvecklade metoden. Detta väcker bland annat frågan om alla mänskliga bias kan reflekteras i statistiska förhållanden i vektormodeller. För att fortsätta undersöka hur vektormodeller kan fånga statistiska förhållanden så utvecklas även en andra metod, WEFAT (*word embedding factual association test*), som låter forskarna vidare undersöka hur ordvektorer kan fånga empirisk information om världen genom en textkorpus. I studien demonstreras metoden genom att låta en vektormodell framgångsrikt återskapa statistiska samband gällande fördelning av kön i olika yrken, samt könsfördelning av androgyna namn.

En studie från 2018 undersöker hur bias och fördomar i en vektormodell förändras i en temporal aspekt (Garg et al, 2018). Som träningsdata för vektormodellen används bland annat en korpus med text som spänner över 100 år. I studien delas korpusen upp i delar som är av längden tio år, och tränar sedan en vektormodell på respektive del. Detta låter forskarna se förändringar hos ord baserat på dess användning i 10-års intervaller. Resultaten av studien är att vektormodellerna tycks fånga viktiga temporala aspekter som exempelvis hur många män och kvinnor som arbetade i olika yrken vid olika tidpunkter, samt hur olika adjektiv och yrken blev mer eller mindre associerade med vissa demografiska grupperingar över tid. I studien visar forskarna att vektormodeller fångar användbara aspekter av den tid som träningsdata kommer från som exempelvis historiska trender och sociala förändringar.

Teori

I detta teoriavsnitt förklaras bakomliggande teorier som krävs för att förstå hur man kan använda resultat från psykologiska tester för att undersöka hur relationer mellan ord i en vektormodell kan anses vara fördomsfulla.

Implicit kognition

Det IAT försöker mäta är kognitiva processer på en implicit, omedveten nivå. En bra utgångspunkt för att beskriva implicit kognition är *system 1* och *system 2*, här beskrivet enligt Kahneman (2013). Enligt 2-systemsperspektivet så agerar alla människor utifrån två distinkta system, där *system 1* är ett omedvetet, automatiskt och snabbt system som aktiveras med liten eller ingen ansträngning alls. *System 2* används till ansträngande intellektuella aktiviteter som kräver direkt uppmärksamhet, som till exempel komplicerade matematiska beräkningar. En människa i *system 2* läge har ofta en subjektiv upplevelse av att man har kontroll, gör medvetna val eller koncentrerar sig. När en människa är i *system 1* läge så sker detta omedvetet, men det innebär inte att människan inte kan utföra komplicerade beräkningar. Kahneman menar att båda systemen används konstant när en människa är vaken, och *system 1* genererar automatiskt och kontinuerligt förslag från intryck, intuitioner, avsikter och känslor som *system 2* överväger. Om *system 2* "godkänner" förslagen så förvandlas, enligt Kahneman, intryck och intuitioner till övertygelser, och impulser förvandlas till viljehandlingar. Upphovsmannen bakom IAT, Anthony Greenwald, delar mycket av den här synen på hur människor fungerar psykologiskt (Greenwald, 2008). För att beskriva de egenskaper som Kahneman associerar med *system 2* så använder sig Greenwald av termen *den första nivån* (eng. *the first level*), och menar att på den första nivån så är tänkandet avsiktligt, rationellt och eftertänksamt. *Den andra nivån* som korresponderar med Kahnemans *system 1*, huserar lägre mentala operationer, som är automatiska, impulsiva och tanklösa och framför allt omedvetna. På den andra nivån, menar Greenwald, kan enkla uppgifter som att knyta skorna, cykla och gå utföras helt automatiskt. Enkla sociala interaktioner som att åka taxi och handla går också att utföra tillsammans med inövade atletiska uppgifter som att spela tennis eller baseball. Vad som också sker på den andra nivån är att associationer skapas från omvärlden. Greenwald menar att världen är full av associationer som människan hela tiden skapar omedvetet mellan objekt. När associationer mellan objekt vä har erhållits så verkar de automatiskt. (Greenwald, 2008).

Omedveten partiskhet och fördomsfullt beteende

De associationer som skapas omedvetet är också de som ligger bakom omedveten partiskhet (eng. *unconscious bias*), vilket innebär att göra bedömningar, fatta beslut baserat på tidigare erfarenhet, djupt rotade tankemönster, antaganden eller tolkningar som inte är medvetna. Detta sker genom att hjärnan automatiskt associerar saker som framträder tillsammans som ihophörande. Således förväntar sig hjärnan omedvetet att dessa saker ska synas tillsammans, med resultatet att när mönster eller kombinationer som inte har samma association dyker upp så känns de onormala och är svårare att processa. Detta leder i bästa fall till neutrala stereotyper och i värsta fall till fördomsfullt och diskriminerande beteende (Frith, 2015).

IAT

Greenwald anser att IAT kan ses som ett fönster som synliggör vissa mentala operationer på en implicit nivå (2008), och syftet med IAT är att mäta styrkan på omedvetna associationer genom att observera svarstider i en kategoriseringsuppgift som administreras genom en dator. Specifikt så undersöks associationer mellan ett koncept och attribut relativt ett annat koncept och attribut genom att mäta skillnader i svarstid vid kategorisering. Traditionellt ges en förklaring för vad man undersöker, men inte hur man undersöker det.

IAT procedur

Proceduren för testet är enligt följande: I ett initialt block av försöksomgångar så visas exempel från två kontrasterande koncept (exempelvis ansiktsbilder av gamla och unga människor) på datorskärmen. Under testets gång visas aldrig mer än ett exempel samtidigt. Försöksdeltagare skall sedan snabbt kategorisera dessa exempel genom att trycka på en av två knappar (exempelvis "e" för gammal och "i" för ung). Deltagarna uppmanas vara så snabba som möjligt genom hela experimentet. Efter ett antal *trials* visas istället exempel från ett par av kontrasterande attribut (till exempel ord som representerar positiv och negativ valens) som också klassificeras med samma knappar som ovan. Steget därefter är en *kombinerad uppgift*, där exempel från samtliga fyra kategorier klassificeras samtidigt och där alla möjliga kategorier är kopplade till samma två knappar (exempelvis "e" för gammal *eller* positiv och "i" för ung *eller* negativ[4]). I en *andra kombinerad uppgift* så används en kompletterande parning av

[4] I originalkällan används andra koncept för att beskriva proceduren som av författaren ansågs problematiska och därför ändrades.

exempel (ex. "e" för ung *eller* positiv och "i" för gammal *eller* negativ). Försöksdeltagare är tvingade att korrigera fel som görs innan de fortsätter, och svarstid mäts fram till nästa korrekta svar. Traditionellt består IAT av sju block, varav fem av dessa räknas som träningsblock och två som rena testblock. Skillnaden i genomsnittlig svarstid mellan de två kombinerade uppgifterna i testblocket utgör grundmätvärdet för IAT. Om en försöksdeltagare har snabbare svarstider för gammal+positiv och ung+negativ än för ung+positiv och gammal+negativ så indikerar det en starkare associativ koppling mellan gammal och positiv valens (Greenwald et al. 2009).

Data och resultat från IAT

Innan insamlad data används så korrigeras den. Detta innefattar bland annat att radera extrema svarstider i båda riktningar (det vill säga för snabba/långsamma), omkoda svarstider utanför förbestämda gränsvärden samt logaritmiskt transformera resultaten. *IAT-effekten* definieras som skillnaden mellan genomsnittlig svarstid mellan de kompatibla och inkompatibla blocken, och effektstorleken beräknades ursprungligen med måttet *d*, där genomsnittliga skillnader i svarstid mellan blocken delas på standardavvikelsen för respektive block. Konventionell liten, mellan och stark effektstorlek är .2, .5 och .8 där ett högre värde innebär en starkare association (Greenwald 1998). Greenwald introducerade 2003 en ny förbättrad variant av effektstorleksberäkningen, som innefattade bland annat att inte systematiskt radera svar med stora skillnader i svarstid. I det nya måttet, *D*-måttet, så används även data från två specifika träningsblock. Vidare så består nämnaren i ekvationen av standardavvikelsen från samtliga testblock, till skillnad från den traditionella uträkningen där standardavvikelsen beräknas per testblock. Det nya *D*-måttet menar Greenwald 1) bättre reflekterar styrkan hos underliggande associationer, 2) på ett kraftfullare sätt bedömer styrkan mellan associationer och andra variabler, 3) förser ökad power för att observera effekten av manipulationer på associationer och 4) tydliggör individuella skillnader som beror på associationer snarare än andra variabler. Med detta menar Greenwald att det nya måttet "kraftigt överträffar den tidigare (konventionella) metoden" (Greenwald, Nosek & Banaji, 2003).

Kritik mot IAT

En av svårigheterna med att bedöma validiteten av IAT enligt Lane, Banaji, Nosek och Greenwald (2007) är att testet representerar ett procedurellt format för att mäta implicit

kognition, snarare än ett mått på ett specifikt konstrukt[5]. Detta innebär att testet är mycket anpassningsbart; IAT-testet kan anpassas till att mäta konstrukt som stereotyper, självförtroende och identitet. Vidare kan IAT-testet användas för att mäta attityder mot koncept som exempelvis kön, etnicitet och favoritmat. Två IAT-tester som mäter samma konstrukt, till exempel attityd mot en etnicitet, kan göra detta med olika stimuli (exempelvis ord och bilder eller olika ord). Detta leder till att två IAT-tester kan ha mycket lite gemensamt, förutom den huvudsakliga strukturen av testet. Test-retest reliabilitet mäter stabiliteten av mätvärden på ett stabilt konstrukt som samlats in från samma person vid två eller fler tillfällen. Vid tester som utförs på individnivå så är ett test-retest värde på 0.90 ett minimum, och på gruppnivå föreslås ett värde på minst 0.70 (Vilagut, 2014). I en analys av 20 studier som använde IAT-testet så varierade test-retest reliabilitet mellan .25 till .69, med ett medel- och medianvärde på .50.

Sedan testet först presenterades 1998 så har det bedrivits mycket forskning inom området implicit social kognition, men trots detta så råder det fortfarande ingen konsensus gällande hur implicita och explicita konstruktioner förhåller sig till varandra. Enligt vissa forskare så är implicit och explicit två helt skilda konstruktioner som inte har något samband alls. Andra forskare menar att dessa konstruktioner har ett samband trots att de är distinkta (se exempelvis Fazio & Olson, 2003; Nosek, 2005).

Vidare har kritik mot måttet *D* lagts fram. Bland annat så menar forskare att det problem som Greenwald försöker lösa med det nya måttet inte enbart är att faktorisera bort individuella skillnader i hastighet mellan försöksdeltagare, utan även tvinga fram en starkare korrelation mellan måtten på implicita och explicita utvärderingar artificiellt.

Oavsett diskussionen kring vad IAT mäter, samt frågor rörande dess validitet och reliabilitet, så har testet fortsatt att användas i främst psykologiska studier, och år 2007 fanns det över 200 studier där testet används för att undersöka fördomar, stereotyper och attityder (Lane et al., 2007). IAT-test har även hittat användningsområden utanför en vetenskaplig kontext. Exempelvis så använd IAT inom marknadsföring, där dess huvudsakliga syfte är att synliggöra omedvetna attityder mot olika varor och varumärken (Perkins, Forehand, Greenwald & Maison 2008).

[5]*Konstrukt* är en term som används i psykologin om begrepp av olika slag utifrån teorin att de är aktivt konstruerade av hjärnans förmåga att sortera och kategorisera.

Distribuerade semantiska modeller

Vektormodeller är en typ av distribuerade semantiska modeller som visar att man kan lära sig mycket om ett ords semantiska innebörd enbart genom att titta på i vilken kontext det förekommer. Den bakomliggande idén för distribuerade semantiska modeller kallas för *distributionshypotesen*, och är ett namn för ett antal antagande gällande språk och menings natur. Enligt distributionshypotesen kommer ord som är semantiskt lika också ha en liknande distribution och orden kommer att förekomma i liknande lingvistiska kontexter (Sahlgren, 2008). Essensen av idén fångas i citatet: "*You shall know a word by the company it keeps*" (JR Firth, 1957 s.11). Vektormodeller lär sig en representation av ords mening baserat på deras förekomst. Mer specifikt så tittar man på *målord* och *kontext*. Ett målord är det specifika ord som man vill undersöka, och kontextord är det eller de ord som förekommer i nära anslutning till målordet. Storleken på kontexten till de ord man vill undersöka kan variera från att vara hela dokument (Dumais, Furnas, Landauer, Deerwester & Harshman, 1988) till bara några få ord (Mikolov, Chen, Corrado & Dean, 2013). För att representera informationen om ordens användning använder modellerna sig av vektorer. Dessa vektorer består av reella tal. Talen eller datapunkterna, som vektorerna består av pekar på olika koordinater inom en vektorrymd med hög dimensionalitet .

Användningsområden

Fördelen med att representera ord med siffror är att det möjliggör geometriska operationer vilket öppnar upp en ny domän av möjliga applikationer. De ord som anses semantiskt lika enligt vektormodellen kan kombineras med andra funktioner för att till exempel hitta olika stavningar på samma ord eller för att undvika stavfel. Operationer kan också utökas för att hantera flera vektorer. Antag att det finns en lista med relaterade ord. Operationer för likhet mellan ord kan då till exempel användas för att hitta fler relaterade ord för att till exempel förlänga listan. Samma typ av operationer kan också användas för att markera de ord i en lista av ord som är minst likt de andra. Operationerna kan också appliceras på dokument, för att ge ett mått på hur lika två eller fler dokument är. Analogier likt de som presenterades i inledningen (*London is to England as Paris is to ...)* visar att algebraiska operationer också går att utföra på vektormodeller med meningsfulla resultat. Denna typ av analogiuppgifter är också en populär metod för att utvärdera vektormodeller, även om det inte är klart vad framgång i analogiuppgifter säger om kvaliteten av ett ords vektor förutom dess förmåga att lösa just denna fråga (Goldberg, 2018). I en rapport från 2019 så påpekar forskarna dock att för att en vektormodell ska kunna uppvisa dessa typer av analogier så krävs det i många fall att parametrar

ändras för att vektormodellen inte ska kunna returnera det närmaste ordet som i många fall är något av de orden som matades in (Nissim, Van Noord & Van der Groot 2019).

Begränsningar med distribuerade semantiska modeller

Distributionshypotesen erbjuder en möjlighet för datorer att kunna representera ords mening enbart baserat på dess förekomst. Yoav Goldberg lyfter dock upp flera begränsningar med ett distributionellt tillvägagångssätt i sin bok "Neural Network Methods for Natural Language Processing" (2018). En av dessa är att det inte går att definiera vad likhet mellan två ord är. Exempelvis för orden *hund, katt* och *tiger* så går det att argumentera för att *hund* och *katt* är mer lika än *katt* och *tiger*, då båda är av typen husdjur. Men det går också att argumentera att *katt* och *tiger* är mer lika, då båda tillhör arten kattdjur. Vilken typ av tolkning som bör användas kan variera i olika tillämpningar och problemet med distribuerade modeller är att de tillåter mycket begränsad kontroll över vilka likheter som skapas.

En annan begränsning med distribuerade modeller enligt Goldberg (2018) är antonymer. Ord som egentligen beskriver motsatser av något (*bra/dålig, köpa/sälja, varm/kall*) tenderar att användas i liknande kontexter. Detta innebär att antonymer hamnar väldigt nära varandra i vektorrymden och anses vara mycket semantiskt lika. Till skillnad från en människa som kan urskilja att varm och kall är motsatser, så kan inte en vektormodell göra denna urskiljning, vilket leder till att vektormodeller inte är lämpade för den typen av tillämpningar som använder sig av antonymer.

När en vektormodell tränas på ord så används ordens kontext för att en semantisk representation av ordet ska skapas. Ironisk nog så är de representationer som skapas i sin tur kontextlösa, vilket innebär att ett ord med två eller flera betydelser inte kan representeras med dessa i vektormodellen. Exempelvis ordet *tomten* kan syfta till både en jultomte men också en gräsplätt; vilket som menas tydliggörs i ordets kontext. *Tomten* får i en vektormodell enbart en representation, som representerar både jultomten och grästomten (Goldberg, 2018).

Corpus bias benämner Goldberg de problem som först identifierades av Bolukbasi et al. (2016) och Caliskan et al. (2017a), och som i förlängningen även denna studie undersöker. Distributionella metoder reflekterar mönster i hur ord används, vilket kan vara på gott och ont. Om uppgiften med vektormodellen är att gissa könet på en karaktär, så kan exempelvis stereotypen att sjuksköterskor ofta är kvinnor och läkare ofta män vara en önskad egenskap. I

många andra tillämpningar så kanske det inte är önskvärt (Goldberg, 2018). Ett typiskt exempel på det här problemet är om man använder Googles översättningstjänst translate och översätter "*she is a doctor, he is a nurse"* till språket turkiska och tillbaka igen så har könen bytts.

Word2vec

Word2vec är en teknik för att skapa vektormodeller som introducerades av forskare på Google (Mikolov et al., 2013). Word2vec använder sig av ett neuralt nät med ett gömt lager för att skapa vektorer. Indata-lagret har lika många neuroner som det finns ord i *vokabulären*[6]. Det gömda lagrets storlek motsvarar vektormodellens tänkta *dimensionalitet* och varierar mellan olika modeller (McCormick, 2019). Dimensionaliteten avser hur många särdrag hos ord som ska lagras, och bestäms empiriskt enligt vad som fungerar bäst givet modellens syfte (Goldberg, 2018). En dimensionalitet på exempelvis 100 innebär att varje ordvektor består av 100 datapunkter som pekar på punkter i den multidimensionella vektorrymden. Utdata-lagret har samma storlek som indata-lagret. Modellen tränas genom att ord och fraser förses till nätverket, som beräknar frekvensen av hur ofta ett målord förekommer i samband med ett annat ord. Vikter i det gömda lagret uppdateras allt eftersom förhållanden mellan specifika ord identifieras. Word2vec kommer i två olika variationer; i skip-gram variationen tränas nätverket att förutspå ett kontextord baserad på ett målord, och i *continuous bag-of-words (CBOW)* variationen tränas nätverket på att förutspå ett målord baserat på kontextord. Denna typ av angreppssätt gör att word2vec-modeller också kallas för prediktiva modeller. När nätverket har tränats så använder man vikterna för respektive ord som representation för ordet; ett ords vektor är vikterna som nätverket använde för att lära sig förutspå ord utifrån kontextorden (McCormick, 2019).

Människan och distribuerad semantik

Det finns de som menar att även människan använder sig av en typ av distribuerad semantik för att förstå vad ord betyder vilket följer av argumentet att när en person säger sig förstå ett ord, så innebär det inte nödvändigtvis att kunna recitera en definition från ett lexikon, utan snarare att personen vet hur man använder det i vardagligt tal. Miller och Charles (1991) menar att för att förstå hur denna kunskap införskaffas måste man först börja med ett antagande; människor lär sig ords betydelse genom att observera hur orden används. Eftersom ord används i fraser och yttranden, så innebär det att denna kontext är av vikt. Fortsatt väljer författarna att kalla den kontextuella information som människor tros använda för att förstå ords innebörd för

[6] Vokabulären innehåller alla möjliga distinkta ord som kan förekomma i vektormodellen.

kontextuella representationer (eng. *contextual representations*) och definierar det som följande; "The *contextual representation* of a word is knowledge of how that word is used" (Miller & Charles, 1991 s.4). Det finns två typer av kontexter, menar Miller och Charles, som kan användas för att lära sig den kontextuella representationen av ett ord. I den smala bemärkelsen så är det enbart de ord som finns direkt före och efter ordet i fråga. Den här inlärningen sker exempelvis vid läsning när en läsare kan skapa sig en förståelse för ett nytt ord. I den breda bemärkelsen så tar man även hänsyn till andra faktorer, som till exempel situationer och intentioner hos de som kommunicerar (Miller & Charles, 1991).

Metod

En stor del av arbetet i denna studie har varit praktiskt. För att kunna utföra studien så har de båda metoderna WEAT och WEFAT programmerats från grunden. I detta kapitel beskrivs därmed de formler som som ligger bakom respektive test. Den färdiga implementationen av båda test återfinns i bilaga 1.

WEAT

Under denna rubrik presenteras de operationer som leder fram till att man kan undersöka hur stark association ett ord har till andra ord i syfte att kunna jämföra detta med IAT-resultat. Samtliga formler bortsett från permutationstestet är hämtade ur Caliskan et al. (2017a; 2017b).

Cosinusmått

Det finns flera metoder för att undersöka förhållanden mellan vektorer. En av dessa är cosinusmåttet (eng. *cosine similarity*). Cosinusmåttet *cos* kan användas för att mäta hur lika två vektorer är. Specifikt för *cos* är att man beräknar vinkeln mellan två vektorer och använder den för att bestämma likheten mellan två vektorers riktning, snarare än till exempel avstånd mellan vektorer i vektorrymden. Givet två vektorer *x, y* där $x = (x_1, x_2 \ldots, x_n)$ och $y = (y_1, y_2 \ldots, y_n)$ så kan cosinusmåttet beräknas med följande formel:

$$cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\Sigma_{i=1}^{n} x_i \cdot y_i}{\sqrt{\Sigma_{i=1}^{n} x_i^2}\ \sqrt{\Sigma_{i=1}^{n} y_i^2}}$$

Formeln ovan beskrivs av Caliskan et al (2017b) som en normaliserad skalärprodukt av två vektorer. Detta innebär att en tidigare mångdimensionell storhet beskrivs med ett tal. Måttet *cos* som produceras av formeln ligger inom ($-1 < cos < 1$), där ett högre värde innebär att vinkeln mellan vektorerna är mindre, vilket innebär att de är mer lika. Lägre värden innebär en större vinkel mellan vektorerna och(Han, Kamber & Pei, 2012).

Associationsstyrka mellan flera ord

Givet att cosinusmåttet mellan vektorer är känt så kan de även användas för att jämföra hur lika två mängder vektorer är i relation till två andra mängder vektorer; antag att det existerar två mängder med koncept (X={advokat, jurist} och {Y=lärare dagispersonal}), samt två mängder med attribut (A={kall, kylig} och B={varm, vänlig}). Nollhypotesen är att det inte existerar någon skillnad mellan de två mängderna av koncept beträffande dess relativa likhet till de två

mängderna av attribut. Låt *w* motsvara ett ospecificerat ord från mängderna med koncept. Formeln för ett ords (*w*) association till de olika orden i *A, B* är

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

där *a* innebär ett specifikt ord i *A* och *b* innebär ett specifikt ord i *B*. Således är *s(w,A,B)* ett mått på hur stark association ett ord, *w*, har i förhållande till samtliga av de två mängderna attribut *A* och *B*. *Mean* innebär att det är det genomsnittliga värdet som avses. Med *cos* avses cosinusmått mellan två vektorer, till exempel mellan *w* och *a*. Med *s(w,A,B)* undersöks styrkan av *w*'s association till orden i båda attribut-mängder genom att subtrahera den genomsnittliga associationsstyrkan till attribut *B* från attribut *A;* om *s(w,A,B)* är ett positivt tal innebär det att *w* har en starkare association till *A*, och om talet är negativt innebär det att *w* har ett starkare association till *B*.

WEAT

Precis som i IAT så är man inte intresserad av ett ords association till de båda attributen, utan man är intresserad av den sammanlagda associationsstyrkan mellan alla ord från mängderna koncept och attribut. Formeln för WEAT är således

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

Om resultatet från testet är ett positivt tal innebär det att koncept *A* har en starkare koppling till *X* relativt *Y*, och om det är ett negativt tal så innebär det en starkare association mellan *Y* och *A* relativt *X*. Detta kan förtydligas med exempel från ovan; Om WEAT resultatet är positivt innebär det att orden *advokat* och *jurist* har en starkare association till orden *kall* och *kylig*, jämfört med orden *lärare* och *dagispersonal*.

Effektstorleken av WEAT

För att beräkna effektstorleken av en undersökning används formeln

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std\_dev}_{w \in X \cup Y} s(w, A, B)}$$

vilket är en mått på hur separerade två distributioner av associationer mellan koncept och attribut är. Måttet, *d*, kan vara positivt, negativt eller 0. Ett positivt tal visar att det finns en starkare association mellan koncept X och attribut A än det finns för koncept Y och attribut A. En negativt tal visar det motsatta och ett resultat på 0 innebär att det inte finns någon skillnad mellan de båda mängderna av koncept i förhållande till attributen.

Permutationstest

Ett permutationstest undersöker i den här studien hur stor sannolikheten är att ett WEAT-resultat är slumpmässigt. I denna studie har det gjorts genom att först observera det riktiga värdet från ett specifikt WEAT med de två listorna med koncept-ord som angetts i respektive orginalstudie. När WEAT-resultatet, $T(x)$, är känt förenar man samtliga ord i koncepten till en ny mängd, $X \cup Y$. Från den nya mängden $X \cup Y$ skapas alla möjliga permutationer[7] av orden i mängden i två lika stora listor. Därefter observeras hur många av dessa permutationer som får ett WEAT-resultat, $T(y)$, som är lika stort eller större än det riktiga värdet från testet. Formeln för permutationstestet är

$$\Pr(\text{exact}) = \sum_{\mathbf{y} \,:\, T(\mathbf{y}) \geq T(\mathbf{x})} \Pr(\mathbf{y})$$

För att beräkna sannolikheten från permutationstestet så delar man antalet gånger en permutation hade ett lika stort eller större värde än det riktiga observerade värdet på det totala antalet permutationer som är möjliga. Detta ger ett värde inom $0 < P \leq 1$. Ett lägre värde innebär att resultaten inte är slumpmässiga. Noterbart är att ett permutationstest aldrig kan resultera i 0, då en möjlig permutation som testas är identisk med orginalutförandet.

WEFAT

WEFAT (*word embedding factual association test*) undersöker vektormodellers förmåga att återspegla empirisk information från verkligheten. Anta att det finns en mängd koncept, som exempelvis yrken, och en faktiskt egenskap i världen, p_w som associeras med varje koncept, till exempel hur många som arbetar inom respektive yrke. WEFAT är utvecklat för att undersöka huruvida ordvektorerna för nämnda koncept har inbäddad kunskap om den associerade egenskapen. Låt *w* stå för ett specifikt koncept ord i en mängd av konceptord, *W* = {doktor, mekaniker, säljare}. Låt *A, B* stå för två lika stora mängder attribut ord, exempelvis (*A* = {han, honom, far} och *B* = {hon, henne, mor}). Egenskapen p_w associerat med varje ord i *W* kan exempelvis stå den faktiska könsfördelning inom ett yrke baserat på en statistisk undersökning. Formeln för att undersöka associationen av ett ord, *w*, till de båda attributen *A, B* är

[7] I WEAT spelar inbördes ordning i permutationerna ingen roll, vilket innebär att den tekniska termen i det här fallet är *kombinationer*.

$$s(w,A,B) = \frac{\text{mean}_{a\in A}\cos(\vec{w},\vec{a}) - \text{mean}_{b\in B}\cos(\vec{w},\vec{b})}{\text{std\_dev}_{x\in A\cup B}\cos(\vec{w},\vec{x})}$$

Resultatet av testet *s(w,A,B)* är ett värde inom -2 och 2, där ett större värde innebär en starkare koppling mellan *w* och *A* relativt *B* och ett mindre värde innebär en stark koppling mellan *w* och *B* relativt *A*. Nollhypotesen är att det inte finns någon koppling mellan *s(w,A,B)* och p_w, vilket undersöks med linjär regression.

Material

I detta avsnitt presenteras det material som har använts vid studiens utförande. Först beskrivs den specifika vektormodell som har använts i studien och därefter beskrivs hur datainsamling har gått till.

Vektormodellen

Vektormodellen som har använts i studien är en prediktiv, 300-dimensionell word2vec modell och har tränats på en korpus bestående av 220 miljoner *tokens*[8]. Samtliga data i korpusen består av nyhetsartiklar från den svenska tidningen Göteborgsposten, och är insamlad mellan åren 2001-2013. Göteborgsposten är Göteborgs största nyhetstidning, och når 60% av Göteborgs befolkning (GP, 2019). En fördel med att använda en korpus från en nyhetstidning är enligt upphovsmakarna att träningsdata är sammanhållande och kurerad (Fallgren, Segeblad & Kuhlmann, 2016). Korpusen i sin helhet är fritt tillgänglig på Språkbankens hemsida (spraakbanken.gu.se). Vokabulären uppgår till ca 192.000 distinkta ord. En begränsning med vektormodellen är att samtliga ord är i gemener. I detta fallet innebär det att fler ord är tvetydiga, exempelvis namnet "Karl" och substantivet "karl", som får samma representation i vektormodellen. Vidare har upphovsmännen till vektormodellen valt att inte inkludera ord som förekommer färre än 25 gånger i korpusen. En utförligare beskrivning av vektormodellen går att finna i Fallgren, Segeblad & Kuhlmann (2016).

Datainsamling

Under denna rubrik beskrivs hur data har samlats in. Först beskrivs hur de IAT-resultat som används för att både skapa och jämföra WEAT-resultat har samlats in. Därefter beskrivs hur data från två statistiska undersökningar har samlats in för att kunna jämföra med WEFAT-resultaten.

IAT-resultat

IAT-resultat från studier där testet har administrerats på svenska, med svenska stimuli, har samlats in digitalt från ett flertal olika källor. Studier där testdeltagarna är barn eller där testet administreras på annat sätt än dator har inte inkluderats. Vidare har IAT-test där annat stimuli

[8] *tokens* innebär ord eller andra betydelsebärande teckensträngar

än ord används inte inkluderats. Utöver vetenskapliga studier så har IAT-resultat även samlats in från Project Implicit (Lofaru, Xu, Nosek & Greenwald, 2018), vilket är en webbsida som tillhandahåller IAT-tester för besökare. Enbart ett av de tester som Project Implicit tillhandahåller uppfyller kraven och inkluderas i denna studie. I studien av Agerström, Carlsson och Rooth (2007) användes ordet *Ameer* i konceptet arabmuslimska män vilket inte återfanns i vektormodellen, och fick strykas i testet. Då både IAT och WEAT kräver att mängderna ord som representerar respektive koncept är av samma längd fick ett ord strykas från konceptet *svenska män*. Vidare saknades attributordet *initiativlös* i vektormodellen och fick strykas tillsammans med ett slumpmässigt utvalt ord från den motsatta attributmängden. I studien av Carlsson & Björklund (2010) saknades attributordet *rättsombud* i vektormodellen vilket innebär att det ströks tillsammans med ett slumpmässigt attributord ur motsatt attributmängd. Slutligen - i Project Implicits undersökning så användes attributordet *ingenjörsvetenskap* som inte finns representerat i vektormodellen. I detta fall ersattes det saknade ordet med *ingenjör*.

Data från statistiska undersökningar

För att undersöka vektormodellens förmåga att representera faktiska egenskaper hos verkligheten så har resultat från två olika statistiska undersökningar samlats in. Det första testet undersöker vektormodellens förmåga att representera män och kvinnors fördelning inom idrott, och det andra testet undersöker vektormodellens förmåga att representera män och kvinnors fördelning inom yrken. Statistik gällande män och kvinnors fördelning inom olika idrotter har samlats in från Riksidrottsförbundets årliga undersökning *RF - Idrotten i siffror* (Idrotten i siffror, 2014). Enbart statistik som visar fördelning mellan *aktiva* kvinnor och män inom respektive idrott har tagits med. Riksidrottsförbundet definierar "aktiv i idrott" som att man deltagit minst en gång per år i någon av föreningens aktiviteter, antingen som ledare, förtroendevald eller aktiv. För studiens ändamål valdes 2014 års statistik ut, då vektormodellen är tränad på data som kommer från en närliggande tidsperiod. Ursprungligen listar Riksidrottsförbundet statistik för 73 olika idrotter, men för studien har enbart de 18 största idrotterna valts[9], baserat på antal aktiva utövare. Motiveringen för att bara välja de största idrotterna är att det ökar sannolikheterna för att idrotten har en bra representation i

[9]Orden som använts för att representera de 18 idrotterna är badminton, bandy, basket, bilsport, bordtennis, kampsport, fotboll, friidrott, golf, handboll, innebandy, ishockey, orientering, ridsport, simning, styrkelyft, tennis och volleyboll.

vektormodellen, då större idrotter bör återkomma med större frekvens i korpusen. En stor idrott innebär i den här rapporten att idrotten hade fler än 25.000 aktiva utövare 2014.

Vidare har idrotter exkluderats om de bryter mot något av följande krav:

1) Idrottens namn består av fler än ett ord (till exempel "amerikansk fotboll")
2) Idrottens namn är tvetydigt (till exempel "cykel", "skidor")
3) Idrottens namn inte finns representerat i vektormodellen (ord som saknades var till exempel "draghund", "flygsport")

Statistik gällande män och kvinnors fördelning inom olika yrken har samlats in från yrkesregistret med yrkesstatistik 2014 (SCB, 2015). Yrkesregistret med yrkesstatistik listar yrkesverksamma inom olika yrken baserat på flera demografiska faktorer. För studiens ändamål valdes en sammanfattning av de 30 största yrkena i Sverige, sett till antalet yrkesverksamma inom respektive yrkeskategori. Då dessa yrkeskategorier kan innefatta flera liknande yrkestitlar så har vissa förändringar behövt göras som kan påverka resultaten. Dessa förändringar har gjorts av författaren till denna rapport. Exempelvis så har yrkeskategorin "Lastbilsförare m.fl." ersatts med "lastbilsförare". I vissa fall har inte yrkeskategorin funnits representerad i vektormodellen, varvid den har ersatts. Exempelvis så har "Lager- och terminalpersonal" ersatts med "lagerarbetare". De yrkeskategorier som ansetts för svåra att representera med ett ord har exkluderats ur studien. Ett exempel på en yrkeskategori som ansågs vara för svår att ersätta med ett ord var "*övriga kontorsassistenter och sekreterare"*, där det kan argumenteras för att en viss könsuppdelning redan existerar i respektive titel och att välja den ena exkluderar den andra. Samtliga av de förändringar som har gjorts härstammar från begränsningar med vektormodellen och dess förmåga att representera ord. Vidare har samtliga förändringar av yrkeskategorier i största mån försökt efterlikna de ursprungliga yrkeskategorier som presenteras av SCB, men där det ej har varit möjligt har en så generell titel som möjligt använts. För en sammanfattning av de förändringar som har gjorts av yrkestitlar se bilaga 2.

Analysen av WEFAT-resultatet som presenteras i denna studie görs i SPSS. Nollhypotesen är att det inte finns något samband mellan WEFAT-resultaten och den faktiska egenskapen från statistiska undersökningar. För undersöka sannolikheten att de observerade sambanden är signifikanta används linjär regression. Effektstorlek av resultaten från WEAT beräknas enligt beskrivning i metodkapitlet och jämförs direkt med respektive IAT studie.

Resultat

Tabell 1 visar en sammanfattning av resultaten från WEAT jämte resultaten från de IAT-studier som har undersökts. Varje resultat jämför två mängder koncept-ord och två mängder attribut-ord. Genomgående har ordlistor från respektive studie används för att replikera IAT-resultatet. N = Antal deltagare; Nk = Antal koncept; Na = Antal attribut. För varje studie rapporteras dess *d* eller *D*-mått, beroende på vad som rapporterats i respektive orginalstudie. För online-IAT (rad 6) har inget *p*-värde rapporterats. *P*-värdet som rapporteras för denna studies undersökning är värdet från permutationstest, vilket inte kan anses vara analogt med *p*-värdet från IAT-studier. Samma resonemang går att applicera på det *d*-mått som rapporteras här, då det ena avser människor och det andra avser ord.

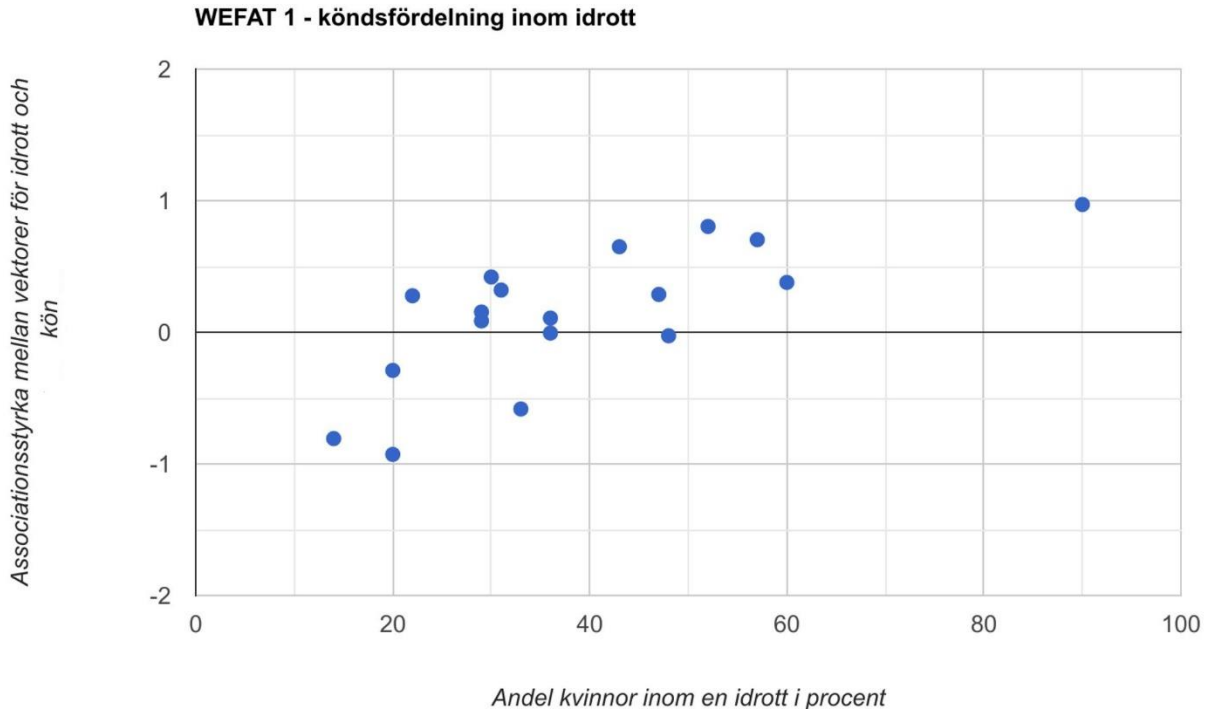
		Orginalresultat					**Resultat från denna studie**			
Koncept	**Attribut**	**Ref[10]**	**N**	***d***	***D***	***P***	**Nk**	**Na**	***d***	***P***
Svenska namn / Arabiska namn	Positiv / Negativ	**1**	157	-	0.64	<0.01	5x2	6x2	1.77	0.0039*
Svenska namn / Arabiska namn	Högpresterande / Lågpresterande	**1**	193	-	0.38	<0.01	5x2	4x2	1.57	0.0039*
Advokat / Förskollärare	Värme / Kyla	**2**	85	-	-0.47	<.01	2x2	2x4	-1.61	0.1667**
Advokat / Förskollärare	Kompetens / Inkompetens	**2**	85	-	0.35	<0.05	2x2	2x4	-0.53	0.5
Manliga namn / Kvinnliga namn	Hem / Karriär	**3**	50	0.85	-	<0.01	7x2	7x2	1.13	0.0143*
Man / Kvinna	Naturvetenskap / Humaniora	**4**	15103	-	0.41	-	8x2	7x2	0.48	0.1720

Tabell 1. Sammanfattning av IAT och WEAT. * = signifikant vid P < 0.05, ** = signifikant vid P < 0.167

Resultaten visar att flera av de associationer som IAT-testet på också går att återskapa i vektormodellen. Rad 1, 2 och 5 visar de studier där signifikant resultat från permuationstester återfanns, och rad 3, 4 och 6 visar de studier där signifikant resultat inte fanns. Noterbart från sammanställningen är att samtliga associationer identifieras i vektormodellen förutom rad 4 (advokat + förskollärare och kompetens + inkompetens), där det motsatta förhållandet identifieras.

I följande stycken beskrivs resultaten från de två WEFAT undersökningarna. Notera att den egenskap förknippad med ett ord som kommer från en statistisk undersökning här förkortas p_w.

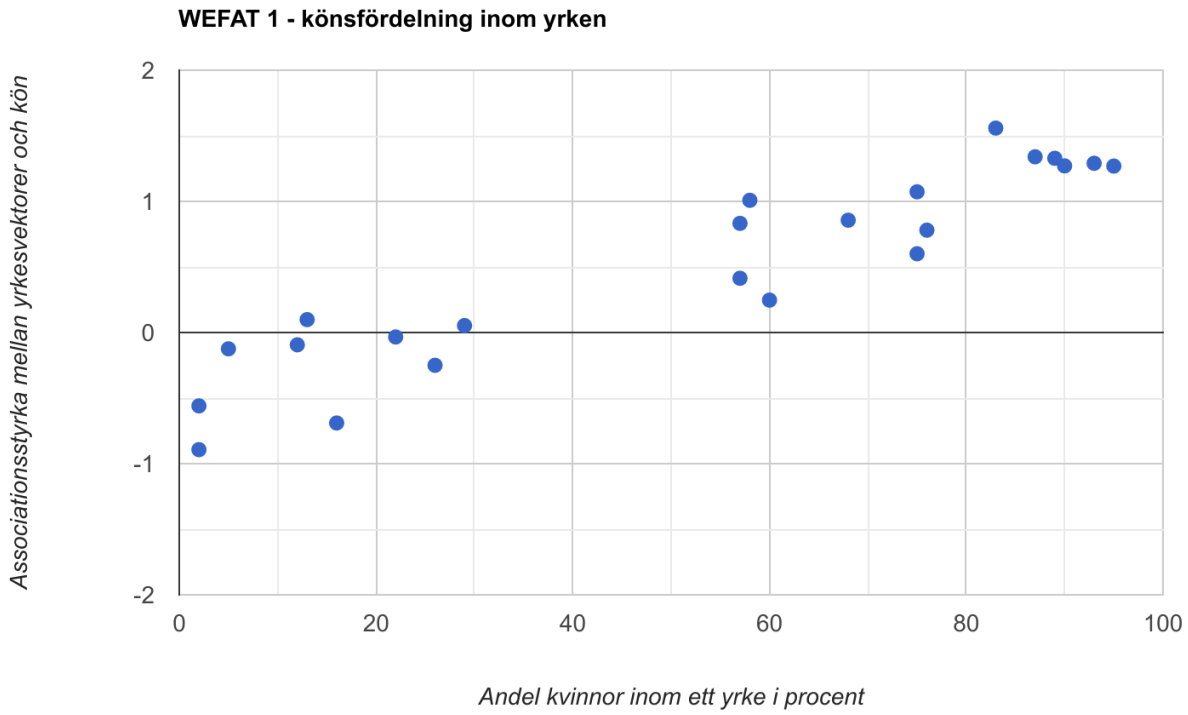
I den första undersökningen är p_w således andel kvinnor i procent som faktiskt utövar en specifik idrott. För att undersöka hur bra WEFAT återspeglade könsfördelningen inom idrott i Sverige gjordes en linjär regressionsanalys för att se om p_w kunde förutspå associationsstyrkan mellan vektorer för kön och idrotter. En signifikant regressionsekvation hittades ($F(1, 16) = 17.612$, $p < .001$), med ett R^2 på .524. Associationsstyrkan för ordvektorer mellan idrott och kön motsvarar $-.664+2.077(p_w)$ associationsstyrka när antalet kvinnor inom en idrott mäts i procent. Associationsstyrkan mellan ord ökar med 2.077 för varje procentenhet. Figur 1 visar resultaten från den första WEFAT-undersökningen, där x-axeln visar andel kvinnor i procent inom en idrott, och y-axeln avser avser associationsstyrkan mellan idrottens namn och ord som representerar kön[10]. En punkt på diagrammet representerar en specifik idrott i figuren



Figur 1. Figuren visar association mellan kön och sport. Pearsons korrelationskoefficient r = .724, p < 0.01.

[10] Orden som används är en direktöversättning av de ord som används i Caliskan et al (2017a) och innefattar *kvinna, tjej, flicka, syster, hon, henne, hennes, dotter* för kvinnligt kön och *man, kille, pojke, bror, han, honom, hans, son* för manligt kön.

För att undersöka hur bra WEFAT återspeglade könsfördelningen inom yrken i Sverige gjordes en linjär regressionsanalys för att se om p_w kunde förutspå associationsstyrkan mellan vektorer för kön och yrken. I denna undersökning är den egenskap, p_w, som är förknippad med ett ord det procentuella antalet kvinnor inom det yrke som ordet representerar. En signifikant regressionsekvation hittades ($F(1, 21) = 139.251$, $p < .000$), med ett R^2 på .869. Associationsstyrkan för ordvektorer mellan yrke och kön motsvarar -.566+2.049 (p_w) associationsstyrka när antalet kvinnor inom ett yrke mäts i procent. Associationsstyrkan mellan ord ökar med 2.049 för varje procentenhet. Figur 2 visar resultaten från den andra WEFAT-undersökningen, där x-axeln visar andel kvinnor i procent inom ett yrke, och y-axeln avser avser associationsstyrkan mellan ordet för yrkets namn och ord som representerar kön. I undersökningen användes samma ord som i föregående undersökning. En punkt på diagrammet representerar ett specifikt yrke i figuren.



Figur 2. Figuren visar association mellan kön och yrke. Pearsons korrelationskoefficient r = .932, P < 0.00.

Diskussion

Fler än hälften av de IAT-resultat som undersöktes med vektormodellen går att finna med signifikanta resultat, vilken kan tolkas som att vektormodellen visar upp tendenser på att associera konceptet arabmuslimska män både med en negativ attityd och en inkompetens-stereotyp relativt svenska män, att män förknippas mer med karriär relativt kvinnor samt att advokater upplevs som kyliga jämfört med förskollärare som upplevs som varma. Dessa associationer som vektormodellen återspeglar visar tydligt inom vilka områden man bör vara försiktig när man använder modellen vid framtida applikationer.

Två resultat härstammar från samma studie där ord fick strykas då de inte fanns representerat i vektormodellen. Detta innebär att konceptet representeras av 2x2 ord, vilket är för lite för att någonsin erhålla signifikant resultat, då minsta möjliga värde som går att erhållas av permutationstestet är $P<0.1666\sim$ för 2x2 ord, vilket också är p-värdet för undersökningen på rad 3 i tabell 1, markerad med **. Detta gör rad 3 till ett specialfall som kommer behandlas som signifikant i resten av denna diskussionen. Resultatet från de två andra undersökningarna tyder på att förhållanden mellan just de orden som kan anses vara fördomsfulla inte existerar i samma utsträckning i denna vektormodellen. Detta gäller i synnerhet resultaten på rad 4, där 50% av permutationer hade samma eller starkare effekt än de ursprungliga koncepten.

Resultaten från de båda WEFAT-testerna visar att vektormodellen fångar empirisk information om vår samtid enbart genom att räkna på ordförekomster i en nyhetstidning. De WEFAT-tester som gjordes i denna studie hade en stark korrelation med fakta från statistiska undersökningar, vilket visar att metoden kan återskapa statistiska egenskaper från verkligheten i en svensk vektormodell.

Sammantaget kan resultaten anses validera de metoder som har använts. Studien har visat att metoderna går att applicera på andra vektormodeller än de som de ursprungligen utvecklades för, för att hitta fördomsfulla associationer som inte tidigare undersökts. Språk tycks inte vara en begränsning för metoderna. Resultaten från studien ger visst stöd för hypotesen att alla mänskliga implicita bias kan finnas reflekterade i statistiska förhållanden hos ord. En svaghet med denna studie är de något undermåliga IAT resultat som använts. Att IAT inte är en lika vedertagen undersökningsmetod i Sverige som i till exempel USA är ganska tydligt då det är betydligt färre studier som har gjorts på svenska där ord har varit den enda typen av stimuli.

Neutrala stereotyper som exempelvis att människor föredrar instrument relativt insekter finns inte på svenska. Flera av de studier som finns på svenska använder även ett relativt litet antal ord för att representera respektive koncept, vilket kan innebära att motsvarande WEAT blir ostabilt jämfört med om man använt flera ord.

Studien ger ett stöd för IAT-testet då man lyckas återskapa vissa av de associationer som ligger bakom bias och stereotyper i en så pass annorlunda miljö. De resultat från permutationstesten som rapporteras i denna studie är inte det samma som de p-värden som rapporteras i respektive orginalstudie, då det förstnämnda avser ord och de sistnämnda avser försökspersoner. Trots detta så visar resultaten från permutationstesten att det inte är en slump att vissa förhållanden mellan koncept och attribut existerar. Oavsett diskussionen om vad det faktiskt är IAT mäter, så kan man även mäta det med viss framgång i en vektormodell. Man kan därför tänka sig det omvända förhållandet, och först undersöka potentiella IAT i en vektormodell för att se om det finns några samband där som skulle kunna motivera en riktig IAT-undersökning.

En insikt från denna studiens arbete är att WEAT-implementationen i kombination med permutationstest inte bara kan användas för att motivera en framtida IAT undersökning, utan även som ledsagning när ett IAT skapas. I kombination med permutationstest kan WEAT på ett snabbt och effektivt sätt visa hur stark effekt ett specifikt ord har jämfört med ett annat genom att undersöka hur P-värdet förändras baserat på ordval. Detta borde vara av intresse för de som undersöker bias med hjälp av ordbaserade IAT, då valet av ord som representerar ett koncept kan anses påverka det konstrukt som undersöks. Denna typ av implementation är mycket snabb, och stora mängder ord kan undersökas med relativt liten ansträngning. Naturligtvis finns dock risken att de specifika koncept man önskar undersöka inte finns representerade i den träningsdata som vektormodellen är tränad på vilket leder till missvisande resultat.

Framtida forskning

I slutändan uppvisar vektormodeller bara de statistiska mönster som redan finns i den textkorpus som används som träningsdata. I denna studie har en textkorpus från en nyhetstidning använts. De implicita associationer som har identifierats i denna studie härstammar således därifrån. Enligt teorin om kontextuella representationer kan vi lära oss om ords betydelse enbart genom att titta på orden som förekommer i nära anslutning. Det vore därför av intresse att skapa och jämföra vektormodeller som exempelvis representerar olika nyhetstidningar med de metoder som använts i denna studie i syfte att diskutera hur olika textkällor kan bidra med implicita

associationer. Att kunna få ett mått på hur fördomsfulla vissa textkällor är kan bidra till ökad kunskap om hur fördomar och stereotyper uppkommer, samt bidra till en diskussion för hur man potentiellt kan minimera de som möjligen härstammar från kontextuella representationer.

I denna studie har enbart två WEFAT-undersökningar gjorts. Båda dessa visade att vektormodellens förmåga att återspegla statistiska egenskaper från verkligheten är god. Båda de WEFAT undersökningar som gjordes är dock mycket lika de undersökningar som gjordes både i Caliskan et al. (2017a) och i Garg et al. (2018), då båda använder sig av könsfördelning som verklig egenskap. Det finns dock ingen anledning att tro att kön är den enda dikotomi som går att undersöka. Det vore därför av intresse att undersöka vilka andra typer av samband som går att effektivt representera i vektormodeller i syfte att bredda användningsområdet men också för att bidra med mer kunskap gällande semantiska vektormodellers förmåga att representera meningsfull information.

Slutsats

Vektormodeller har en kuslig förmåga att kunna återspegla empirisk fakta om sakers beskaffenhet enbart genom att räkna på vilka ord som används tillsammans. Denna studie har undersökt två relativt nya metoder som går att applicera på vektormodeller för att extrahera ny typ av information. Metoderna går att applicera med relativt gott resultat på redan existerande modeller för att belysa ordförhållanden som kan leda till fördomsfulla applikationer, men bör också gå att använda i syfte att lära oss mer om fördomar. Att använda vektormodeller för att representera faktiska egenskaper baserad på ordfrekvenser är en ny typ av tillämpning som bör vidare undersökas.

Källor:

Agerström, J., Carlsson, R., & Rooth, D-O. (2007). Etnicitet och övervikt: implicita arbetsrelaterade fördomar i Sverige. *Institutet för arbetsmarknadspolitisk utvärdering; Vol. 2007:19.* Institutet för arbetsmarknadspolitisk utvärdering (IFAU)

Bermudez, J., L. (2014). *Cognitive Science - An introduction to the Science of the Mind*. 2 uppl. Cambridge: Cambridge University Press

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Debiasing Word Embedding. In *30th Conference on Neural Information Processing Systems*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017a). Semantics derived automatically from language corpora contain human-like biases. *Science*. https://doi.org/10.1126/science.aal4230

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017b). Supplementary Materials for: Semantics derived automatically from language corpora contain human-like biases. *Science*. https://doi.org/10.1126/science.aal4230

Carlsson, R., & Björklund, F. (2010). Implicit stereotype content : Mixed stereotypes can be measured with the Implicit Association Test. *Social Psychology*. https://doi.org/10.1027/1864-9335/a000029

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa., P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88*. https://doi.org/10.1145/57167.57214

Fallgren, P., Segeblad, J., & Kuhlmann, M. (2016) Towards a standard dataset of Swedish word vectors. *In Proceedings of the Sixth Swedish Language Technology Conference (SLTC)*, Umeå, Sweden, 2016.

Fazio, R. H., & Olson, M. A. (2003). Implicit Measures in Social Cognition Research: Their Meaning and Use. *Annual Review of Psychology*. https://doi.org/10.1146/annurev.psych.54.101601.145225

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *The Philological Society*.

Frith, U. (2015). *Unconscious bias*. The Royal Society. Hämtat från royalsociety.org/-/media/policy/Publications/2015/unconscious-bias-briefing-2015.pdf

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. arXiv preprint arXiv:1711.08412.

Geniesse, M., & Taylor, M. (2019). Implicit vs. explicit testing: Identifying what consumer really believe. Hämtad från https://www.quirks.com/articles/implicit-vs-explicit-testing-identifying-what-consumers-really-believe

Goldberg, Y. (2018). Neural network methods for natural language processing. *Computational Linguistics*. https://doi.org/10.1162/COLI_r_00312

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G. (2008, May). The Psychology of Blink: Understanding how our minds work unconsciously - Part 1 of 2. *Public lecture (Allen L. Edwards Psychology Lectures).* Podcast tillgänglig på UWTV: http://www.uwtv.org/programs/displayevent.aspx?rid=24708

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/a0015575

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Idrotten i siffror. (2014) *Riksidrottsförbundet.* Hämtad från https://www.rf.se/Statistik/Idrottenisiffror

Tosik, M., Lygteskov Hansen, C., Goossen, G., & Rotaru, M. (2015). Word Embeddings vs Word Types for Sequence Labeling: the Curious Case of CV Parsing. https://doi.org/10.3115/v1/w15-1517

Irsoy, O., & Cardie, C. (2014). Deep recursive neural networks for compositionality in language. In *Proceedings of The Neural Information Processing Systems*.

Kahneman, D. (2013). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kullinger, J, (2016). Övertalande budskap gör implicita könsrollsattityder mer traditionella och explicita könsrollsattityder mer egalitära. *Mid Sweden University, Department of Psychology.* http://urn.kb.se/resolve?urn=urn:nbn:se:miun:diva-27189

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and Using the Implicit Association Test: IV: What We Know (So Far) about the Method. In *Implicit measures of attitude*.

Lofaro, N., Xu, F., Nosek, B., & Greenwald, A. G. (2018) Gender IAT Sweden [dataset]. *Open Science Framework*, Tillgängligt hos: https://osf.io/p6gx2/.

McCormick, C (2019). *The Inner Workings of word2vec*. Publisher: Author

Mikolov, T., Chen, K., Corrado, C., & Dean, J,. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781[csCL] Hämtad från https://arxiv.org/abs/1301.3781

Miller, G. A., & Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*. https://doi.org/10.1080/01690969108406936

Nalisnick, E., Mitra, B., Craswell, N., & Caruana, R. (2017). Improving Document Ranking with Dual Word Embeddings. https://doi.org/10.1145/2872518.2889361

Nissim, M., Van Noord, R. & Van der Groot, R,. (2019). Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor. **arXiv:1905.09866 [cs.CL]** Hämtad från https://arxiv.org/abs/1905.09866

Perkins, A., Forehand, M., Greenwald, A. G., and Maison, D. (2008). The influence of implicit social cognition on consumer behavior: measuring the non-conscious. *Handbook of Consumer Psychology*, eds C. Haugtvedt, P. Herr, and F. Kardes (Hillsdale, NJ: Lawrence Erlbaum Associates), 461–475.

Bias. (n.d.). In *Psykologiguiden.se*, Retrieved from https://www.psykologiguiden.se/psykologilexikon/?Lookup=bias

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*.

SCB (2015) Yrkesregistret med yrkesstatistik 2014. Hämtad från https://www.rf.se/globalassets/riksidrottsforbundet/dokument/statistik/rf_i_siffror_2014.pdf

Vilagut, G. (2014) Test-Retest Reliability. *Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research*. Springer, Dordrecht https://doi.org/10.1007/978-94-007-0753-5

Weizenbaum, J. (1966). ELIZA---a computer program for the study of natural language communication between man and machine. *Communications of the ACM*. https://doi.org/10.1145/365153.365168

Winograd, T. (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. Applied Mathematics.

Bilaga 1

Pythonkod WEAT, WEFAT och grundläggande vektoroperationer. Samtlig kod är skapad för denna studies ändamål av författaren till denna rapport.

```
import gensim, logging, math, operator, pickle, itertools, statistics
"""
Program can be used to calculate WEAT and WEFAT from txt-vectors
along with basic vector-operations such as similarity.
AUTHOR: Michael Jonasson micjo469@student.liu.se
INSTRUCTIONS:Replace "path_to_vector" with real path to txt vector"""
on last rows, replace [koncept], [koncept], [attribut], [attribut]
with lists of concepts and attributes. Run from terminal.
"""

def list_from_vec():
    """Made specifically to split and make list of txt-vectors"""
    new_list = []
    open_file = open(path_to_vector)
    file_content = open_file.read().split('\n')  # split on newline
    open_file.close()
    for line in file_content:
        split_line = line.split()  # split on emptyspace
        for ele in split_line[1:]:
            float_converted = float(ele)
            split_line.append(float_converted)  # add floats and
            split_line.remove(ele)  # removes str ele
        new_list.append(split_line)

    new_list = new_list[:-1]  # removes list index out of range
    return(new_list)

def two_word_similarity(word1, word2):
    """legacy Function computes cosine similarity of two words"""
    w1 = find_word(word1)
    w2 = find_word(word2)
    numerator = 0  # whats above in dot product formula
    denom_b = 0  #      ...below in ...
    denom_a = 0  #      ...below in ...
    i = 1
    for num in w1[1:]:
        multiply = num * w2[i]      # a * b
        denom_a += num * num          # a^2
        denom_b += w2[i] * w2[i]  # b^2
        i += 1
        numerator += multiply
    b = math.sqrt(denom_b)  # square of denom b
    a = math.sqrt(denom_a)  # square of denom a
    cosine = numerator / (a * b)
    return(cosine)
```

```
def most_similar(word, numb):
    """Function list most similar words"""
    words = list_from_vec()
    word = find_word(word)
    word_cosine = {}
    for line in words:
        if line[0] == word[0]:
            pass
        else:
            numerator = 0  # whats above in dot product formula
            denom_b = 0  #      ...below in ...
            denom_a = 0  #      ...below in ...
            i = 1
            for num in line[1:]:
                multiply = num * word[i]      # a * b
                denom_a += num * num          # a^2
                denom_b += word[i] * word[i]  # b^2
                i += 1
                numerator += multiply
            b = math.sqrt(denom_b)  # square of denom b
            a = math.sqrt(denom_a)  # square of denom a
            cosine = numerator / (a * b)
# a * b / ((sqrt of a) * (sqrt of b)) = cosine similarity
            word_cosine.update({line[0] : cosine})
    sorted_vector  =  sorted(word_cosine.items(),  key=operator.itemgetter(1),  reverse=True)
             # sorts vector based on distance
    for val in sorted_vector[:numb]:  # numb is number of key:val pairs
        print (val)

def find_word(arg):
    """Fetch wordvector corresponding to inputword """
    words = list_from_vec()
    c = ""
    for line in words:
        if arg == line[0]:
            c = line
        else:
            pass
    return c

def cosine_similarity(w1, w2):
    """ Calculates and returns cosine similarity of two words """
    numerator = 0  # whats above in dot product formula
    denom_b = 0  #      ...below in ...
    denom_a = 0  #      ...below in ...
    i = 1
    for ele in w1[1:]:
        multiply = ele * w2[i]      # a * b
        denom_a += ele * ele          # a^2
        denom_b += w2[i] * w2[i]  # b^2
```

```
        i += 1
        numerator += multiply
    b = math.sqrt(denom_b)  # square of denom b
    a = math.sqrt(denom_a)  # square of denom a
    cosine = numerator / (a * b)  # a * b / ((sqrt of a) * (sqrt of b)) = cosine similarity
    return(cosine)

def mean_cosine(target, attr_a, attr_b):
    """ Calculate mean cosine for target - used in WEAT"""
    cos = 0
    for a in attr_a:
        cos += cosine_similarity(target, a)
    mean_cos_a = cos / len(attr_a)
    cos = 0
    for b in attr_b:
        cos += cosine_similarity(target, b)
    mean_cos_b = cos / len(attr_b)
    swab = (mean_cos_a - mean_cos_b)
    return(swab) # returns swab

def main(t1, t2, a1, a2):
    """The main function for WEAT-test """
    words = list_from_vec()  # splits words from file
    target_list1 = find_multiple(t1, words)  # collect vectors for t1
    target_list2 = find_multiple(t2, words)
    attribute_list1 = find_multiple(a1, words)
    attribute_list2 = find_multiple(a2, words)
weat    =    word_embedding_association_test(target_list1,    target_list2,    attribute_list1,
attribute_list2)
    combinations = create_permutation(t1+t2)
    pval = p_value(combinations, attribute_list1, attribute_list2, words, weat)
    return(weat)

def word_embedding_association_test(t1, t2, a1, a2):
    """WEAT as described in Caliskan - returns swab"""
    sxab = 0
    syab = 0
    for t in t1:
        sxab += mean_cosine(t, a1, a2)
    for z in t2:
        syab += mean_cosine(z, a1, a2)
    mean = (sxab + syab) / len(t1 * 2)
    stdev = std_dev(t1, t2, a1, a2, mean)
    sxab = sxab / len(t1)
    syab = syab / len(t2)
    weat = (sxab - syab) / stdev
    return(weat)

def word_embedding_association_test_non_perm(t1, t2, a1, a2):
    """standalone WEAT without permutationtest for fast weats"""
    words = list_from_vec()  # splits words from file
```

```
    target_list1 = find_multiple(t1, words)  # collect vectors for t1
    target_list2 = find_multiple(t2, words)
    attribute_list1 = find_multiple(a1, words)
    attribute_list2 = find_multiple(a2, words)
    sxab = 0
    syab = 0
    for t in target_list1:
        sxab += mean_cosine(t, attribute_list1, attribute_list2)
    for z in target_list2:
        syab += mean_cosine(z, attribute_list1, attribute_list2)
    mean = (sxab + syab) / len(target_list1 * 2)
    stdev = std_dev(target_list1, target_list2, attribute_list1, attribute_list2, mean)
    sxab1 = sxab / len(target_list1)
    syab1 = syab / len(target_list2)
    final = (sxab1 - syab1) / stdev
    return(final)

def p_value(combinations, attribute1, attribute2, words, weat):
    """Calculates P-value from permutation/combinations"""
    c = 0
    right = len(combinations) -1
    for i in range(len(combinations)):
        t1 = find_multiple(combinations[i], words)
        t2 = find_multiple(combinations[right], words)
        #print(combinations[i], combinations[right])
        right -= 1
        score = word_embedding_association_test(t1, t2, attribute1, attribute2)
        if score >= weat:
            #print(combinations[i+1], combinations[right])
            #print(score, ' is larger than ', weat)
            c += 1
        print(i)
    print('number of combinations: ', len(combinations))
    print('number of bigger: ', c)
    print('pvalue: ', c / len(combinations))

def word_embedding_factual_association_test_new(t1, a1, a2):
    """WEFAT as described by caliskan"""
    words = list_from_vec()  # splits words from file
    target_list = find_multiple(t1, words)  # collect vectors for t1
    attribute_list1 = find_multiple(a1, words)
    attribute_list2 = find_multiple(a2, words)
    all_attributes = attribute_list1 + attribute_list2
    wefats = {}
    for t in target_list:
        list_for_stan_dev = []
        swab = mean_cosine(t, attribute_list1, attribute_list2)
        for a in all_attributes:
            cos = cosine_similarity(t, a)
            list_for_stan_dev.append(cos)
        st_dev = statistics.stdev(list_for_stan_dev)
```

```
        new = swab / st_dev
        wefats.update({t[0] : new})
    sorted_dict = sorted(wefats.items(), key=operator.itemgetter(1), reverse=True)
    for key in sorted_dict:
        print(key)

def std_dev(target_list1, target_list2, attribute_list1, attribute_list2, mean):
    """Calculate standard dev based of targets, attributes and mean"""
    all_targets = []
    sq_sum =0
    all_targets = target_list1 + target_list2   #now in union
    a_list = []
    for target in all_targets:
        test = mean_cosine(target, attribute_list1, attribute_list2)
        a_list.append(test)
        swab = mean_cosine(target, attribute_list1, attribute_list2)
        distance = swab - mean
        square = distance * distance
        sq_sum += square
    sq_sum2 = sq_sum / (len(all_targets)-1)  # Bessel's correction
    stand = statistics.stdev(a_list)
    stdev = math.sqrt(sq_sum2)
    return stand

def find_multiple(words, wordlist):
    """Find vectors of list of words"""
    num = len(words)
    found = []
    c = 0
    for target in words:
        for w in wordlist:
            if target == w[0]:
                found.append(w)
                c += 1
    return(found)

def create_permutation(t):
    """Create all possible combinations from a list of words"""
    r = int(len(t)/2)
    c = 0
    combinations = []
    for combo in itertools.combinations(t, r):
        combination = []
        for com in combo:
            combination.append(com)
            c+=1
            print(c)
        combinations.append(combination)
    return combinations

if __name__ == "__main__":
```

```
  print(main([koncept], [koncept], [attribut], [attribut]))  # returns WEAT score
  print(word_embedding_factual_association_test_new([koncept][koncept]  # return WEFAT
[attribut][attribut]))
```

Bilaga 2 - Yrkeskategorier

I vänsterspalten i följande tabell visas ursprungliga kategorier från "Yrkesregistret med yrkesstatistik" från 2014 (SCB, 2015). I högerspalten visas de omskrivningar som gjordes för att anpassa yrkeskategorier till vektormodellen.

Undersköterskor, hemtjänst, hemsjukvård och äldreboende	undersköterska*
Grundskollärare	grundskollärare
Företagssäljare	säljare
Butikssäljare, fackhandel	butiksbiträde*
Barnskötare	barnskötare
Vårdbiträden	vårdbiträde
Lager- och terminalpersonal	lagerarbetare
Förskollärare	förskollärare
Butikssäljare, dagligvaror	butiksbiträde*
Städare	städare
Mjukvare- och systemutvecklare m.fl	systemutvecklare
Restaurang- och köksbiträden	restaurangpersonal
Lastbilsförare	lastbilsförare
Vårdare, boendestödjare	vårdare
Maskinställare och maskinoperatörer, metallarbete	maskinoperatör
Grundutbildade sjuksköterskor	sjuksköterska
Träarbetare, snickare m.fl	snickare
Ekonomiassistent m.fl	ekonomiassistent
Undersköterskor, vård och specialavdelning	undersköterska*
Säljande butikschefer och avdelningschefer i butik	butikschef
Installations- och serviceelektriker	elektriker
Gymnasielärare	gymnasielärare
Civilingenjörsyrken inom elektroteknik	elingenjör
Banktjänsteman	banktjänsteman

* avser ord som förekommer flera gånger, vilket var möjligt då den ursprungliga könsfördelningen var snarlik i båda kategorier. Kategorier som stryks från undersökningen då lämpligt ersättningsord inte identifierades var "*övriga kontorsassistenter och sekreterare*", "*Planerare och utredare m.fl*", "*Kockar och kallskänkor*".