עיבוד שפה טבעית ש13-14:
סמנטיקה של מילים;
שיכוני מילים (embeddings)
SLP 6 & G, E 14

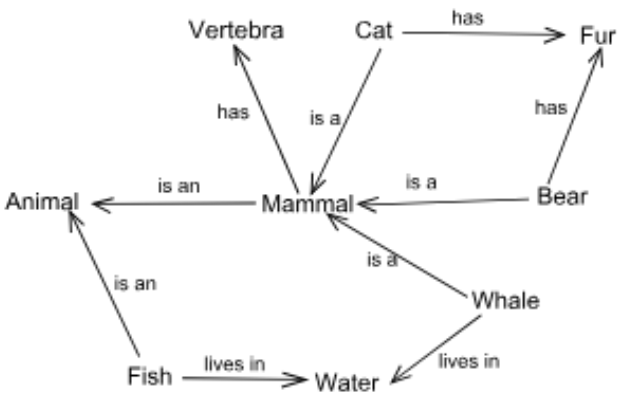
דרכים לתת משמעות למילים

- הגדרה מילונית
- קשר למילים אחרות
  - *אלון* הוא סוג של *עץ* שהוא סוג של *צמח* (is-a)
  - *לעץ* יש *עלים* (has)
  - *שמחה* היא ההפך מעצב
  - *מכירה* היא ההיפוך התפקידי של *רכישה*; הפועל *רכש* מתאר *רכישה*
- (גזירה היסטורית?)

קשרים בין מילים ומשמעויות (words and senses)

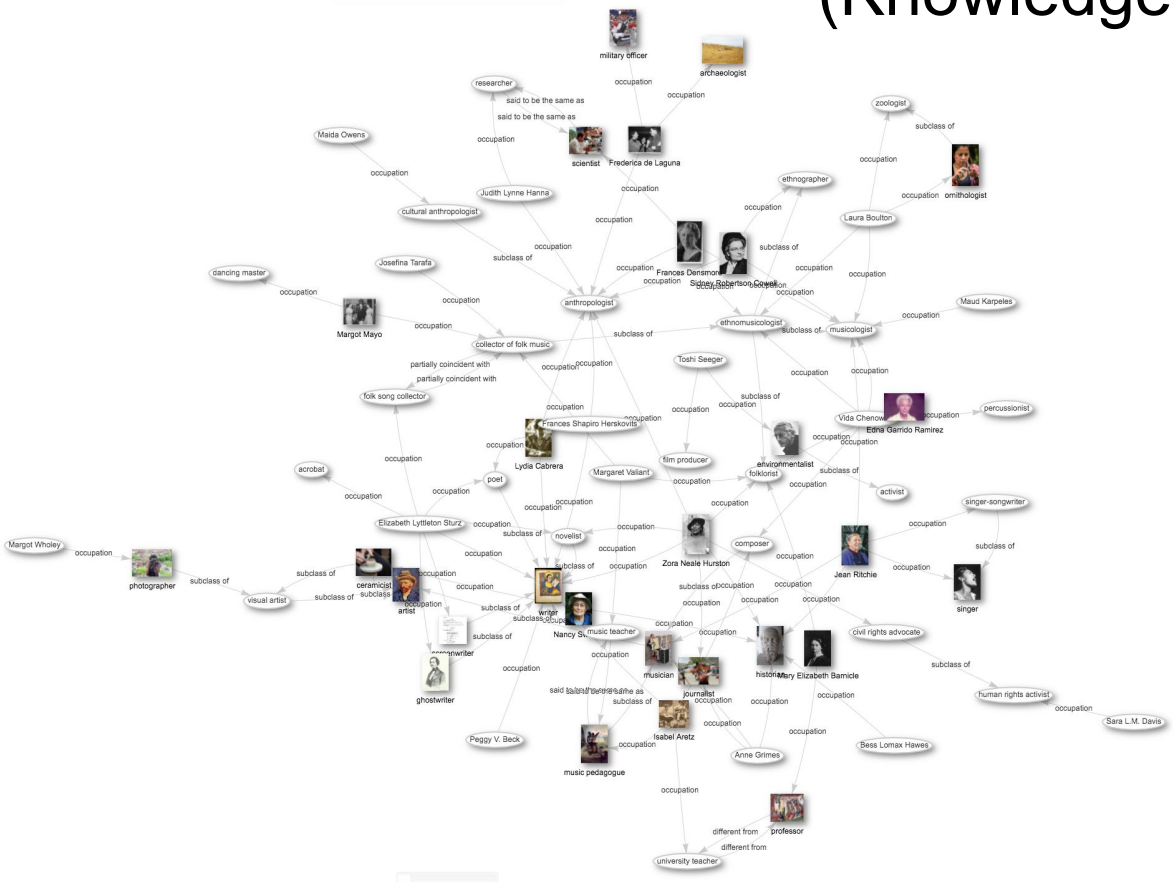
- מילה אחת, הרבה משמעויות: **פוליסמיה** (polysemy)
  - כוכב; עכבר
- כמה מילים, אותה משמעות: מילים **נרדפות** (synonymy)
  - ירח / לבנה / סהר
- מילים **דומות** (similarity): מתארות חפצים או רעיונות דומים בעולם
  - כלב / חתול; עצב / קנאה
- מילים **קשורות** (relatedness): מתארות חפצים או רעיונות קשורים לאותו **שדה סמנטי**
  - כלב / מלונה; מורה / לוח; עצב / דמעה

רשתות סמנטיות (Semantic Nets)



- משאב שנוצר ידנית, מעין "מילון מרושת"
- מילים (lemmas) ממופות למובנים (senses), והמובנים מקושרים ביניהם בצורת גרף
- המשאב הכי מפורסם, הכי בשימוש, עם הכי הרבה נגזרות: WordNet
  - נגזרת לתמונות: ImageNet
  - לשפות שאינן אנגלית: BabelNet
  - למונחים מה"עולם האמיתי": ConceptNet
  - למידע על תפקידי מילים במשפט: VerbNet, FrameNet

קשור: גרף ידע (Knowledge Graph)



ConceptNet

Related terms

- en branch
- en squirrel
- en wood
- en plant
- en nest
- en shade
- en leaf
- en leaves
- en stick
- en branches
- en paper
- en apple
- en big
- en forest
- en plant
- en trunk
- en wood
- en tall
- en climb
- en oak

More »

Things located at tree

- en a bird
- en a leaf
- en fruit
- en a snake
- en a squirrel
- en a acorn
- en some leaves
- en roots
- en earth
- en sap
- en shade
- en wood
- en a bald eagle
- en a dead leaf
- en fallen leaves
- en grass
- en the ground
- en a marmot
- en moss
- en a root

More »

Types of tree

- en Something you find outside
- en b tree
- en cherry tree
- en r tree
- en aalii (n. plant)
- en acacia (n. plant)
- en African walnut (n. plant)
- en albizzia (n. plant)
- en alder (n. plant)
- en angelim (n. plant)
- en angiospermous tree (n. plant)
- en anise tree (n. plant)
- en apple tree
- en arbor (n. plant)
- en aroeira blanca (n. plant)
- en ash (n. plant)
- en Australian nettle (n. plant)
- en avocado tree
- en balata (n. plant)
- en banyan tree

More »

Synonyms

- pt Árvores (n. plant)
- ja 木 (n)
- ar شَجَرَة (n. plant)
- ca arbre (n. plant)
- ca arbre (n. shape)
- ca diagrama arbori (n. shape)
- da træ (n. plant)
- da trævækst (n. plant)
- eo arbo
- sh drvo
- sh stablo
- en Sir Herbert Beerbohm Tree (n. person)
- en shoetree (v. change)
- en corner (v. motion)

tree is capable of...

- en grow
- en grow leaves
- en shade the lawn
- en branch out
- en branch into leaves
- en cast a shadow
- en drop leaves
- en shade the car
- en shade from the sun
- en branch in several directions
- en branch off and get bigger

tree is used for...

- en shade
- en beauty
- en climbing
- en wood
- en a wind break
- en add beauty
- en aesthetics
- en bird nests
- en breaking wind
- en build a tree house in
- en building a treehouse

tree is a type of...

- en plant
- en a tall plant
- en woody plant (n. plant)
- en plane figure (n. shape)
- en a living thing
- en usually about twenty feet tall
- en wood
- de plant
- en diagram
- en figure
- en fixed obstacle (n)

Symbols of tree

- mul
- mul
- mul
- mul
- mul

tree is a way of...

- en chase (v. motion)
- en plant (v. contact)

tree is made of...

- en wood
- en leaves

Properties of tree

- en alive

Things located near tree

- en A leaf
- en a squirrel

Things made of tree

- en talbe
- en wood

tree wants...

- en sunshine
- en water

Etymological roots of "tree"

- ang treo
- enm tre

Antonyms

- en bush
- en shrub
- en seed
- en paper
- en squirrel
- en plant
- en palm
- en nice

Etymologically derived terms

- ja ツリー

Things motivated by tree

- en make it grow

tree is created by...

- en planting a seed

Derived from

- ko 뿔 (n)

tree doesn't want...

- en be cut down

Distinct terms

- en specific

https://conceptnet.io/c/en/tree

רשתות סמנטיות

- תחזוקה של רשת סמנטית היא **המון** עבודה
- לא תמיד אופן התיוג מוסכם על המשתמשים או אפילו על המתייגים
- אם רוצים סוג קשר חדש, צריך לעבור על כל המילון מחדש
- הרחבה לתחום / שפה חדשים - **המון** עבודה ולכן כמעט תמיד (חצי-)אוטומטי ולכן תמיד כולל שגיאות
  - קונספטנט עברית: מקשרת בין **עץ** ל**חרש (שם עצם)** שמקושר מיד ל**כבד-שמיעה**

הגישה ההקשרית / התפלגותית Distributional Semantics

- "דע את המילה לפי חברותיה" You shall know a word by the company it keeps
  - פירת' Firth 1957
- הגישה טוענת כי אפשר להגדיר מילה רק באמצעות אוסף ה**הקשרים** בהם היא מופיעה
  - "תשתמש.י במילה הזאת במשפט" → אמרו לכמן פעם? אמרתקם למישהו.י פעם?
- גישה **מאוד** ידידותית לשיטות חישוביות:
  - ההקשרים ניתנים לחילוץ מקורפוס גדול
  - לא צריך לתייג כלום, לא צריך לאצור מילונים
  - (באיזו בעיה אנחנו מנסים למצוא מילה לפי ההקשר שלה?)
- מהו "הקשר"?

התופעות הלקסיקליות בראי ההקשר

- מילה אחת, הרבה משמעויות: **פוליסמיה** (polysemy)
  - כוכב; עכבר
- כמה מילים, אותה משמעות: מילים **נרדפות** (synonymy)
  - ירח / לבנה / סהר
- מילים **דומות** (similarity): מתארות חפצים או רעיונות דומים בעולם
  - כלב / חתול; עצב / קנאה
- מילים **קשורות** (relatedness): מתארות חפצים או רעיונות קשורים לאותו **שדה סמנטי**
  - כלב / מלונה; מורה / לוח; עצב / טיפול

מילה אחת, הקשרים שונים

כמה מילים, אותם הקשרים

כמה מילים, הקשרים "דומים"

מילים בהקשר זו של זו

דקה על מורפולוגיה

- הבחנו בין הטיה (inflection) לגזירה (derivation)
- הטיה, כמו בכל מילון, לא מקבלת יחס ("עץ" ו"עצים" שולחים לאותו ערך, באמצעות **למטיזציה** lemmatization)
- מה עושים עם גזירה? האם ליצור קשר מפורש בין *יֶלֶד* ו*ללדת*?
  - גישת וורדנט: **כן**, אבל עם קשר מיוחד (derivationally-related form)
- מה לגבי תכוניות שאינן מוטות אבל משפיעות על הטיה, כמו *מין דקדוקי*?

היתרון העצום של מיפוי מילים למרחב (= שיכון)

כשמשתמשים בשיכונים כשכבת קלט לבעיה
(כמו סיווג מסמך), לא חייבים להסתמך על הופעה
של מילים **ספציפיות** בסט האימון -
ידע שנלמד על מילים מסוימות **מועבר באופן טבעי**
למילים שקרובות להן במרחב הווקטורי

מילים משוכנות (embedded) במרחב שמימדיו הם תכוניות (properties)

- ערכיות valence: עד כמה המונח "נעים"
- גירוי arousal: מה עוצמת התחושה הנובעת מהמונח
- דומיננטיות dominance: כמה שליטה מגולמת במונח
- ...

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

תופעות לקסיקליות בראי ההקשר והמרחב הווקטורי

- מילה אחת, הרבה משמעויות: **פוליסמיה** (polysemy)
  - כוכב; עכבר
- כמה מילים, אותה משמעות: מילים **נרדפות** (synonymy)
  - ירח / לבנה / סהר
- מילים **דומות** (similarity): מתארות חפצים או רעיונות דומים בעולם
  - כלב / חתול; עצב / קנאה
- מילים **קשורות** (relatedness): מתארות חפצים או רעיונות קשורים לאותו **שדה סמנטי**
  - כלב / מלונה; מורה / לוח; עצב / טיפול

מילה אחת, הקשרים שונים

כמה מילים, אותם הקשרים

כמה מילים, הקשרים "דומים"

מילים בהקשר זו של זו

וקטור שהוא ממוצע של וקטורים רחוקים זה מזה

וקטורים קרובים זה לזה

וקטורים קרובים עם קשר אלגברי צפוי

וקטורים עם קואורדינטות מסוימות דומות

מטריקות?

- איך נדע אם אוסף שיכונים באמת מייצג דמיון לשוני אמין?
- דרך אחת: לשאול אנשים עד כמה מילים הן דומות
  - סובייקטיבי מאוד
  - לא מתייחס לתכוניות השונות
  - מה עם הקשר?

word1	word2	POS	SimLex999	conc(w1)	conc(w2)	concQ	Assoc(USF)	SimAssoc333	SD(SimLex)
old	new	A	1.58	2.72	2.81	2	7.25	1	0.41
smart	intelligent	A	9.2	1.75	2.46	1	7.11	1	0.67
hard	difficult	A	8.77	3.76	2.21	2	5.94	1	1.19
happy	cheerful	A	9.55	2.56	2.34	1	5.85	1	2.18
hard	easy	A	0.95	3.76	2.07	2	5.82	1	0.93
fast	rapid	A	8.75	3.32	3.07	2	5.66	1	1.68
happy	glad	A	9.17	2.56	2.36	1	5.49	1	1.59
short	long	A	1.23	3.61	3.18	2	5.36	1	1.58
stupid	dumb	A	9.58	1.75	2.36	1	5.26	1	1.48
weird	strange	A	8.93	1.59	1.86	1	4.26	1	1.3

https://fh295.github.io/simlex.html

Two words are *synonyms* if they have very similar meanings. Synonyms represent the same *type* or *category* of thing. Here are some examples of synonym pairs:

- *cup / mug*
- *glasses / spectacles*
- *envy / jealousy*

In practice, word pairs that are not exactly synonymous may still be very *similar*. Here are some very similar pairs - we could say they are nearly synonyms:

- *alligator / crocodile*
- *love / affection*
- *frog / toad*

In contrast, although the following word pairs are *related,* they are not not very similar. The words represent entirely different types of thing:

- *car / tyre*
- *car / motorway*
- *car / crash*

In this survey, you are asked to compare word pairs and to rate how *similar* they are by moving a slider. Remember, things that are related are not necessarily similar.

If you are ever unsure, think back to the examples of synonymous pairs (*glasses / spectacles*), and consider how close the words are (or are not) to being synonymous.

There is no right answer to these questions. It is perfectly reasonable to use your intuition or gut feeling as a native English speaker, especially when you are asked to rate word pairs that you think are not similar at all.

מטריקות?

- איך נדע אם אוסף שיכונים באמת מייצג דמיון לשוני אמין?
- דרך אחת: לשאול אנשים עד כמה מילים הן דומות
  - סובייקטיבי מאוד
  - לא מתייחס לתכוניות השונות
  - מה עם הקשר?
- ומהצד השני - מהי הגדרת דמיון בתוך המרחב הווקטורי?
  - דמיון קוסינוס Cosine similarity: מכפלה פנימית מנורמלת
  - מרחק קוסינוס (אם נרצה פונקציית הפסד דווקא): אחת פחות הדמיון
  - (מה הטווחים?)

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

חילוץ וקטורי מילים מתוך הקשרים

- תיוג תכונות של כל מילות העולם זה מפרך וסובייקטיבי
- במקום זה אפשר להשתמש בווקטורי ייצוג מחולצים מקורפוס גדול - ההקשרים נותנים את המימדים

מטריצת מילים / מסמכים

- נייצג כל מילה לפי אוסף המסמכים שהיא מופיעה בו (וכמה בכל מסמך)
- האם מרחק קוסינוס ייתן לנו אומדן טוב למרחק סמנטי כלשהו?
- חסרון מרכזי?
- בפועל, טכניקה מאוד מקובלת ב**אחזור מידע** (חיפוש / Information Retrieval)

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

	וואלה	הארץ	YNET	ישראל היום
הזוי	22,800	5,300	74,400	1,690
נתניהו	99,300	668,000	122,000	35,100
"יצחק תשובה"	9,870	9,600	13,100	2,330
"שמעון אדף"	223	664	555	57

מטריצת שכנויות

- נספור הופעות של זוגות מילים באותו מסמך / "הקשר"
- אינפורמציה הדדית אי-שלילית (PPMI)

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

Figure 6.10 Co-occurrence counts for four words in 5 contexts in the Wikipedia corpus, together with the marginals, pretending for the purpose of this calculation that no other words/contexts matter.

	p(w,context)					p(w)
	computer	**data**	**result**	**pie**	**sugar**	**p(w)**
cherry	0.0002	0.0007	0.0008	0.0377	0.0021	0.0415
strawberry	0.0000	0.0000	0.0001	0.0051	0.0016	0.0068
digital	0.1425	0.1436	0.0073	0.0004	0.0003	0.2942
information	0.2838	0.3399	0.0323	0.0004	0.0011	0.6575
p(context)	0.4265	0.4842	0.0404	0.0437	0.0052	

Figure 6.11 Replacing the counts in Fig. 6.6 with joint probabilities, showing the marginals around the outside.

מטריצת שכנויות

- נספור הופעות של זוגות מילים באותו מסמך / "הקשר"
- אינפורמציה הדדית אי-שלילית (PPMI)
- חסרון מרכזי - המטריצה הזאת עצומה (V x V)
- נתחיל למצוא דרכים לשכן ב**מרחב קטן** (יחסית)

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

TF-IDF

- גישה מקובלת מאוד עדיין באחזור מידע (Information Retrieval, IR) הידוע ביישומו המרכזי **מנוע חיפוש**
- מטרה: להתאים מילים מהשאילתא (query) למסמך (document) מתוך האינדקס (index)
- נסיון ראשון: ניתן ציון לכל מסמך לפי כמה פעמים מופיע בו כל מונח (term) מהשאילתא
  - זה "שכיחות מונחים", term frequency, בקיצור tf
- בעיה מיידית: מילות יחס ושאר מיליות ישתלטו על הציון
  - פתרון מיידי: נעיף אותן מהשאילתא (**stop words**)
- בעיה מיידית 2: מילות תוכן נפוצות עדיין ישתלטו
- פתרון: נמשקל כנגד ה-tf את ה**ייחודיות** של מילות השאילתא באינדקס: ככל שמילה מופיעה בפחות מסמכים, כך היא תהיה יותר חזקה
  - זה inverse document frequency, או בקיצור idf
- נתבל הכל במעט אבקת לוג כדי להימנע מהשפעות-יתר של חריגים

Number of documents in index

$$\boxed{\text{tf-idf}_1(t,d)=\log(1+f_{t,d})*\log(N/n_t)}$$

term frequency

Number of documents with the term t

שיכון ע"י חיזוי הקשרים מקומי - Skip-Gram

- הוצע ב-2013 יחד עם שיטה נוספת בשם CBOW (שק-מילים הקשרי), ביחד נקראות word2vec
- הרעיון הכללי: נפעיל אלגוריתם **חיזוי**, אבל נשמור רק את המשקלות הנלמדים

- בהינתן מילה w בקורפוס, ננסה לחזות את השכנות שלה c (בהקשר) בהסתברות גבוהה ככל הניתן, ונעבור כך על כל המילים בקורפוס $P(c|w)$
  - בקרה עצמית (self-supervision)
- יעד החיזוי - **רגרסיה לוגיסטית** $P(c|w) \propto \sigma(emb(w) \cdot emb(c))$
- בעצם נחשב $P(+|w,c)$ (״מה ההסתברות ש-w ו-c מתרחשות ביחד״)
  - לכל מילה יהיה וקטור **מטרה** (target) בתפקידה כחוזה, ווקטור **הקשר** (context) בתפקידה כנחזית
    - שתי מטריצות! (בסוף בד״כ זורקים את מטריצת ההקשרים)
- נרצה למקסם תחזיות של מילים אמיתיות בהקשר ושל **דגימות אקראיות** (Negative Sampling) מתוך אוצר המילים $P(+|w,c_{pos}) \prod (P(-|w,c_{neg}))$

- (היפר-פרמטרים?)

דוגמה

- משפט: **אכלתי אתמול עוגה טעימה מאוד**.
- מילת היעד: **עוגה**
- רוחב החלון: 2, דגימות שליליות: 3
- דוגמאות חיוביות: (עוגה, אכלתי), (עוגה, אתמול), (עוגה, טעימה), (עוגה, מאוד)

  c_{pos} w
- כמה דוגמאות שליליות? (עוגה, אקספוזיציה), (עוגה, אפרכסת), (עוגה, טבת), ...

  c_{neg} w

סקיפגראם - פרטים

הפסד:

$$
\begin{aligned}
L_{CE} &= -\log\left[P(+|w,c_{pos})\prod_{i=1}^{k}P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos}) + \sum_{i=1}^{k}\log P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos}) + \sum_{i=1}^{k}\log\left(1-P(+|w,c_{neg_i})\right)\right] \\
&= -\left[\log \sigma(c_{pos}\cdot w) + \sum_{i=1}^{k}\log \sigma(-c_{neg_i}\cdot w)\right]
\end{aligned}
$$

גרדיאנט:

$$
\begin{aligned}
\frac{\partial L_{CE}}{\partial c_{pos}} &= [\sigma(\mathbf{c}_{pos}\cdot\mathbf{w})-1]\mathbf{w} \\
\frac{\partial L_{CE}}{\partial c_{neg}} &= [\sigma(\mathbf{c}_{neg}\cdot\mathbf{w})]\mathbf{w} \\
\frac{\partial L_{CE}}{\partial w} &= [\sigma(\mathbf{c}_{pos}\cdot\mathbf{w})-1]\mathbf{c}_{pos} + \sum_{i=1}^{k}[\sigma(\mathbf{c}_{\mathrm{neg}_i}\cdot\mathbf{w})]\mathbf{c}_{neg_i}
\end{aligned}
$$

עדכון:

$$
\begin{aligned}
\mathbf{c}_{pos}^{t+1} &= \mathbf{c}_{pos}^{t} - \eta[\sigma(\mathbf{c}_{pos}^{t}\cdot\mathbf{w}^{t})-1]\mathbf{w}^{t} \\
\mathbf{c}_{neg}^{t+1} &= \mathbf{c}_{neg}^{t} - \eta[\sigma(\mathbf{c}_{neg}^{t}\cdot\mathbf{w}^{t})]\mathbf{w}^{t} \\
\mathbf{w}^{t+1} &= \mathbf{w}^{t} - \eta\left[[\sigma(\mathbf{c}_{pos}\cdot\mathbf{w}^{t})-1]\mathbf{c}_{pos} + \sum_{i=1}^{k}[\sigma(\mathbf{c}_{\mathrm{neg}_i}\cdot\mathbf{w}^{t})]\mathbf{c}_{neg_i}\right]
\end{aligned}
$$

דגימות שליליות

- דגימה אחידה גורמת למילים נדירות להיות מיוצגות כמעט רק בהקשרים שליליים (כי?)
- לדגום יותר מדי פעמים את המילים הנפוצות בקורפוס זה "לא כל-כך מעניין"
- נדגום מתוך התפלגות מוסחת לטובת מילים נדירות:

$$P_\alpha(c) = \frac{count(c)^\alpha}{\sum_c count(c)^\alpha}$$

ערך אמפירי טוב לאלפא = 0.75 (עובד גם עם PPMI, ולא במקרה)

מאפיינים עיקריים של SkipGram

- המימד קטן במידה שנוחה לנו (50-2000 הוא טווח מקובל)
- גבולות המסמכים לא חשובים כ״כ
- רוחב החלון כן חשוב. באנגלית:
  - חלון קטן (2) - מקרב מילים בעלות קשר תחבירי (Hogwarts ~ Sunnydale, Evernight)
  - חלון גדול (+5) - מקרב מילים בעלות קרבה סמנטית / נושאית (Hogwarts ~ Dumbledore, Malfoy)
- במטריצת מסמכים, מילות מבנה (function words) מוסיפות רעש ולרוב נתעלם מהן (stopwords). ב-skipgram הן חשובות

- האם סקיפגראם זה מידול שפה?

- מה לגבי pre-processing?

- (עוד וריאנטים של SG - חלון דינמי, משקל לפי מרחק, דילוג מעל מטרות)

מאיפה הדאטא

- ויקיפדיה - dumps חודשיים, זמינים לכל, הרבה עבודת נקיון
- קורפוס ווסטברי (Westbury) - ויקיפדיה אנגלית "נקייה" מ-2010
- הזחילה הגדולה (CommonCrawl) - מאות טרה-בית של דפים מהרשת
  - עוד דומים: C4, הערימה The Pile
- ~~קורפוס ספרים (BooksCorpus)~~ - הושג בניגוד לזכויות יוצרים
- רשתות חברתיות
- ועוד ועוד ועוד

מאיפה הדאטא

- טוויטר עברית:

https://u.cs.biu.ac.il/~yogo/tw2v/similar

מילים דומות. בטוויטר. בעברית.

קראנו הרבה ציוצים והרצנו אלגוריתם שלומד איך מילים מתנהגות.
הכניסו מילה ותקבלו את המילים שהמחשב חושב שדומות לה, בסדר יורד.

תפוח חפש

דימיון	מילה	שכיחות
100.0	תפוח	15744
76.4	ותפוח	608
73.7	תפוחי	6733
72.8	אבוקדו	7885
70.8	שומשום	1860
66.6	דג	16882
68.5	תפוז	5253
68.1	מלפפון	5577
68.1	זיתים	5170
66.2	בורקס	13798

אפל חפש

דימיון	מילה	שכיחות
100.0	אפל	33302
91.0	סמסונג	14405
89.7	מיקרוסופט	11085
86.2	אמזון	6278
78.9	גוגל	48585
83.4	וואווי	2055
80.0	בלקברי	2535
80.0	אסוס	1760
79.9	קוואלקום	1237
79.2	אינטל	6360

גוריון חפש

דימיון	מילה	שכיחות
100.0	גוריון	29891
79.8	גורין	50
74.1	גוריונית	26
72.7	נון	4083
69.7	ארי	21695
72.8	רודס	732
70.3	כספית	12919
71.9	גריון	67
70.9	עזאי	56
69.1	גביר	4893

ריגשתה חפש

דימיון	מילה	שכיחות
100.0	ריגשתה	101
65.6	התקלת	271
64.0	הפתעתם	210
64.1	ביאסתם	107
59.0	ריגשת	3130
63.9	כפרהעליך	75
63.3	י'מלך	35
62.6	שוקראן	238
62.3	שוקרן	228

גישושים בדאטא

סמנטעל

"לא משחקת, זה מתיש אותי"

- ענת קם, *Twitter*

המילה של אתמול הייתה **גלידה**. היום, חידה מספר **288**, ציון הקרבה של המילה הכי קרובה (999/1000) למילה הסודית היום הוא **79.58**, ציון הקרבה של המילה העשירית הכי קרובה (990/1000) הוא 75.19 וציון הקרבה של המילה האלף הכי קרובה (1/1000) הוא 62.07.

ניחוש | ניחוש

#	ניחוש	קרבה	מתחמם?
10	מלונה	35.97	(רחוק)
8	חתול	46.93	(רחוק)
7	חול	38.84	(רחוק)
6	חלון	31.44	(רחוק)
9	גמל	25.43	(רחוק)
4	תורה	24.17	(רחוק)
1	עיבוד	21.48	(רחוק)
2	יהושע	13.49	(רחוק)
5	עבודה	7.32	(רחוק)
3	משחק	6.58	(רחוק)

נחשו את המילה הסודית

אפשר לנחש מילה או ביטוי קצר. המילה הסודית יכולה להיות בכל אורך, אבל תהיה מילה בודדת (ולא ביטוי).

המשחק יגיד כמה המילה קרובה סמנטית למילה הסודית. קרבה לא באה לידי ביטוי באיות אלא במשמעות. "משמעות" או "קרבה סמנטית" נמדדת באמצעות Word2Vec. או במילים פשוטות יותר: שתי מילים הן יותר קרובות סמנטית ככל שיותר סביר להשתמש בשתיהן בקונטקסט דומה. (ספציפית בויקיפדיה, כי זה הקורפוס שהמודל אומן עליו) הציון לקרבה הסמנטית הוא בין 100- ל־100, כש־100- זה ממש רחוק ו־100 זאת המילה הסודית בעצמה.

המילה הסודית יכולה להיות כל חלק דיבר, אבל תמיד תהיה מילה בודדת.

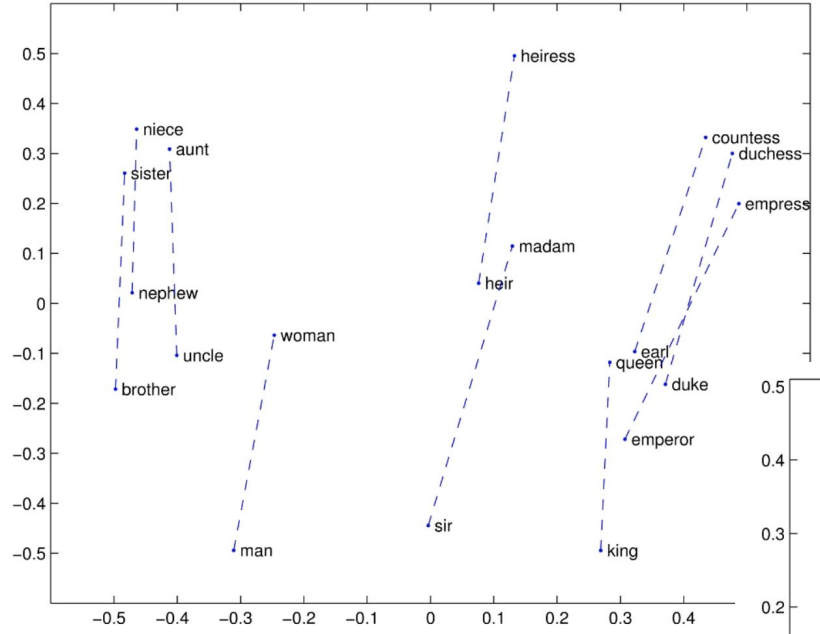
המשחק יגיד כמה הניחוש שלך קרוב למילה הסודית ויציין אם המילה אחת מ1000 המילים הכי קרובות למילה הסודית.

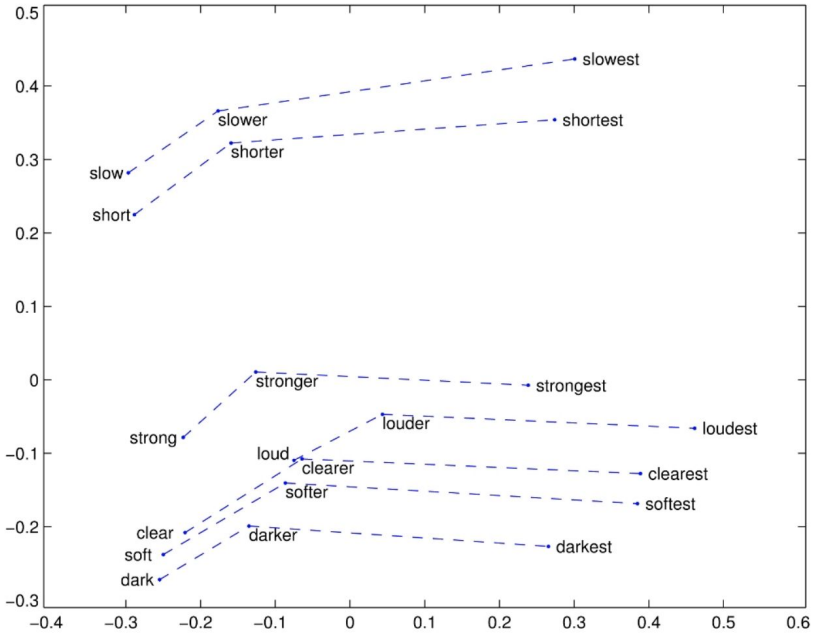
בניגוד ל־Wordle, בדרך כלל צריך יותר מ6 ניחושים. יותר בכיוון של כמה עשרות. ויש מילה חדשה כל יום.

(איתמר שפי)

שיכונים והמרחב הווקטורי

- מילים דומות וקשורות (ראינו)
- אנלוגיות

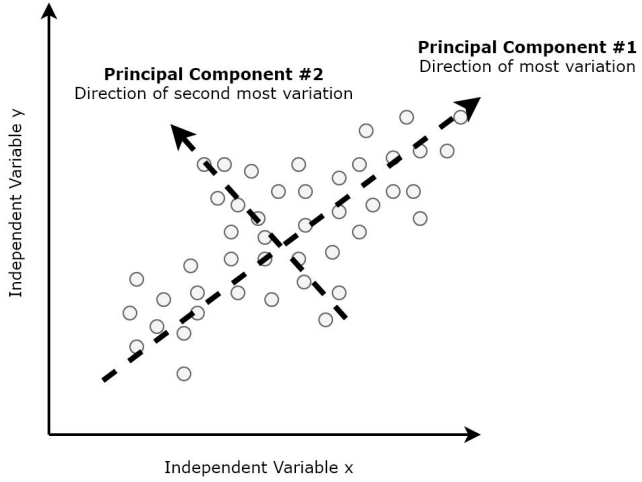




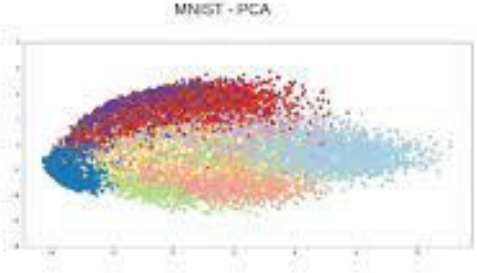
שיכונים והמרחב הווקטורי

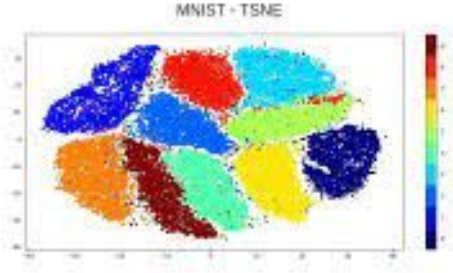
- מילים דומות וקשורות (ראינו)
- אנלוגיות
- ניתוחי שכנים קרובים nearest neighbors (ראינו)
- הרחבות - מילים עם כמה משמעויות?
- מהו הקשר? אולי ניתוח תלויות יכול לעזור?

הצגה במימד נמוך



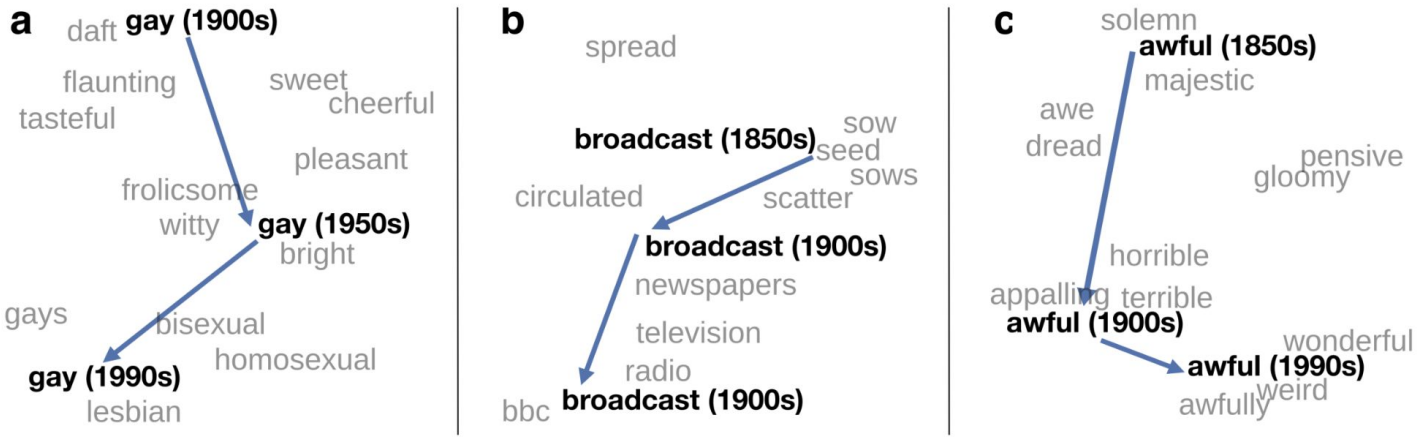
- ניתוח מרכיבים עיקריים (PCA - Principal Component Analysis)
  - מוצא קורלציה בין הנקודות
  - משמר **שונות**
- שיכון שכנים סטוכאסטי לפי התפלגות טי (t-SNE)
  - אלגוריתם **לומד** (לפי דאטא נתון)
  - מקרב התפלגויות של **מרחקים** בין זוגות נקודות - במימד גבוה ובמימד נמוך
  - (יקר חישובית - ניתן להתחיל עם PCA)





שיכונים דיאכרוניים

- לוקחים קורפוס מתקופות שונות
- רואים איך משמעויות משתנות - **סחף סמנטי**



חסרונות של מודלי שיכון מילים

- מילים שמחוץ לאוצר המילים (Out-of-vocabulary - **OOV**)
  - (6 סוגים לפחות?)
  - (3 סוגי התמודדות לפחות)
  - (שקף אחרון)
- ייצוג ביטויים (New York)
- קושי ב**הסברנות** (מה מתאר כל מימד בשיכון?)
- מקורם 100% מהדאטא - פתח ללימוד **הטיות** בלתי-רצויות
  - מלך - ילד + ילדה = מלכה
  - רופא - ילד + ילדה = ?

משימת דמיון מילים - מטריקה בפועל (המשך)

- אמרנו קודם שאין משמעות למיקומים של המילים במרחב עצמו, או לסקאלה שבה דמיונות המילים מחושבים לעומת זו שמתויגת ע״י בני אדם

- לכן נשווה בין **דירוגים** של אוסף שיפוטי דמיון בין זוגות, ונשתמש במטריקות של **קורלציה** correlation
  - קלט: שתי רשימות מדורגות של הזוגות הנשפטים
  - פלט: מספר בטווח [1, 1-] שמתאר את הקשר שבין הרשימות

מטריקות קורלציה

- מטריקות לשערוך מודלי דמיון:
- Pearson correlation

מטריקות קורלציה

- מטריקות לשערוך מודלי דמיון:
- Pearson correlation

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

מטריקות קורלציה

- מטריקות לשערוך מודלי דמיון:
- Pearson correlation

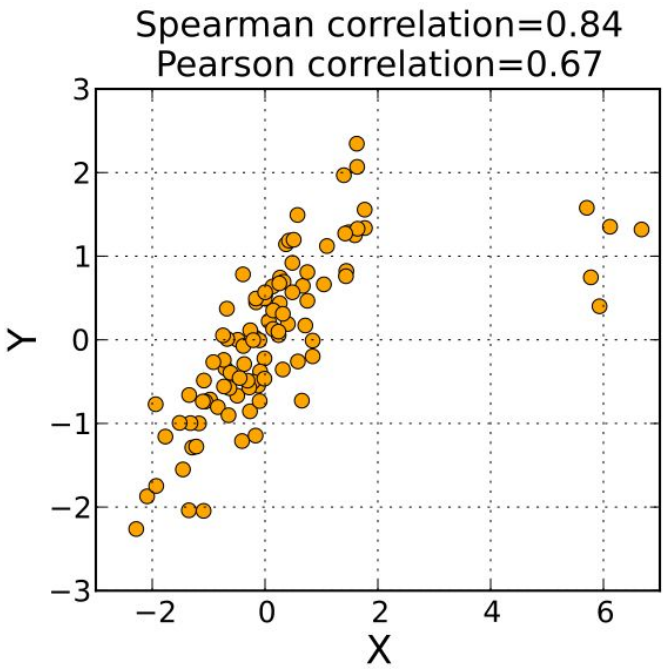
$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

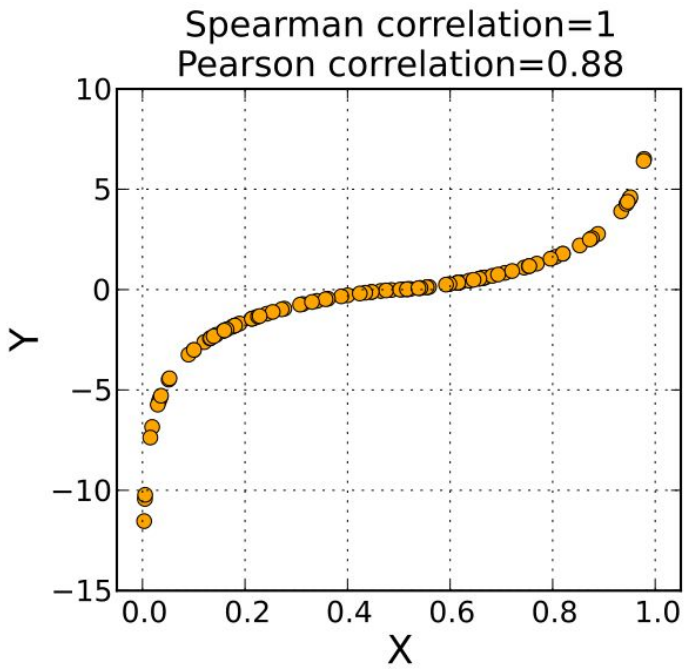
- Spearman correlation
  - R - דירוג (rank)
  - נוסחא (2) נכונה לדירוגים ללא שוויון (d_i הוא ההפרש עבור נקודה i)

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X),\mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}}$$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

אינטואיציה להבדל בין פירסון לספירמן





בעיית ה-OOV

- אאוט אוף ווקאביולרי - מילים שלא מופיעות בטבלת השיכונים
  - 7 סוגים של מילים כאלה
    - ישויות חדשות
    - מילה מאוד מאוד נדירה
    - תאריכים ומספרים
    - מילים מדומיינים מיוחדים
    - תחדיש
    - מילים משפות זרות, מילים שאולות
    - שגיאטת דפסס
  - 4 דרכים לפתרון
    - [הכנסה אקטיבית של שגיאות לקורפוס אימון]
    - תיקון שגיאות ב-inference
    - בחירת קורפוס אימון מגוון
    - מידול תווים / תת-מילים