Phrase2vec in Practice

Aerin Kim

BYOR

At the end of this talk

```
AERINs-MacBook-Pro:bitbucket_BYOR aerin$ python WTB_phrase_similarity.py
Loading the data file... Please wait...
Successfully loaded 3.6 G bin file!
##########################################################################
##########################################################################
########## WELCOME TO THE PHRASE SIMILARITY CALCULATOR ##########
##########################################################################
##########################################################################
Type the phrase1: How Trump Would Stimulate the U.S. Economy
Type the phrase2: Trump unveils plan to revitalize America's economy
##########################################################################
Similarity Score:  0.732695
##########################################################################
Type the phrase1: Let's make America great again
Type the phrase2: We will make America prosperous and powerful
##########################################################################
Similarity Score:  0.670172
##########################################################################
Type the phrase1: Let's make America great again
Type the phrase2: We are going to build a beautiful wall
##########################################################################
Similarity Score:  0.438079
##########################################################################
```

Download & Install

- https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit
- easy_install -U gensim
- easy_install numpy
- easy_install scipy
- Pip install nltk

Frustration

- Binary Vector (discrete representation)
- Apple = [0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]

AND

- Fruit = [0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0]

= 0

- Dimensionality? – very sparse representation

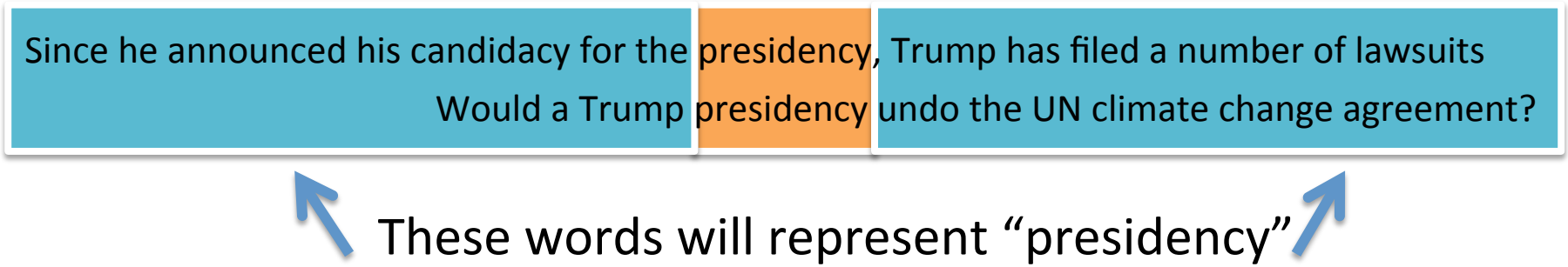
Solution

- Statistical NLP ➔ Co-occurrence matrix

	'buildings'	'factories'	'has'	'owns'	'Donald'	'Trump'	'many'	'hands'	'in'	'Mr.'	'.'	'China'	'small'	'relatively'
'buildings'	0	0	0	0	0	0	1	0	0	0	1	0	0	0
'factories'	0	0	0	0	0	0	1	0	1	0	0	0	0	0
'has'	0	0	0	0	0	2	1	0	0	0	0	0	0	1
'owns'	0	0	0	0	0	1	1	0	0	0	0	0	0	0
'Donald'	0	0	0	0	0	2	0	0	0	0	0	0	0	0
'Trump'	0	0	2	1	2	0	0	0	0	1	0	0	0	0
'many'	1	1	1	1	0	0	0	0	0	0	0	0	0	0
'hands'	0	0	0	0	0	0	0	0	0	0	1	0	1	0
'in'	0	1	0	0	0	0	0	0	0	0	0	1	0	0
'Mr.'	0	0	0	0	0	1	0	0	0	0	0	0	0	0
'.'	1	0	0	0	0	0	0	1	0	0	0	1	0	0
'China'	0	0	0	0	0	0	0	0	1	0	1	0	0	0
'small'	0	0	0	0	0	0	0	1	0	0	0	0	0	1
'relatively'	0	0	1	0	0	0	0	0	0	0	0	0	1	0

Better Solution

- Predict surrounding words of every word
- You shall be judged by the company you keep



Word Embedding (Word2Vec)

- Objective function:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-m\leq j\leq m, j\neq 0} \log p(w_{t+j}|w_t)$$

MAXIMIZE the log porbability of any context word given the current center word.

One (very) Big Vector Θ

- Θ is the set of ALL parameters in one vector

$$\theta = \begin{bmatrix} v_{aardvark} \\ v_a \\ \vdots \\ v_{zebra} \\ u_{aardvark} \\ u_a \\ \vdots \\ u_{zebra} \end{bmatrix}$$

OBJ function of single window

$$p(o|c) = \frac{\exp\left(u_o^T v_c\right)}{\sum_{w=1}^{W} \exp\left(u_w^T v_c\right)}$$

- Trump announced his candidacy for president as a Republican

U V U

- Assuming window size = 1
- First element: Exp(U^T(his) $\cdot$ V (candidacy))
- Second element: Exp(U^T(for) $\cdot$ V (candidacy))

Result after the Optimization

- Semantically

Famous example:

Vec(King) – Vec(man) = Vec(Queen) – Vec(woman)

Vec(CSS) – Vec(Front-end) = Vec(Django) – Vec(Beck-end)

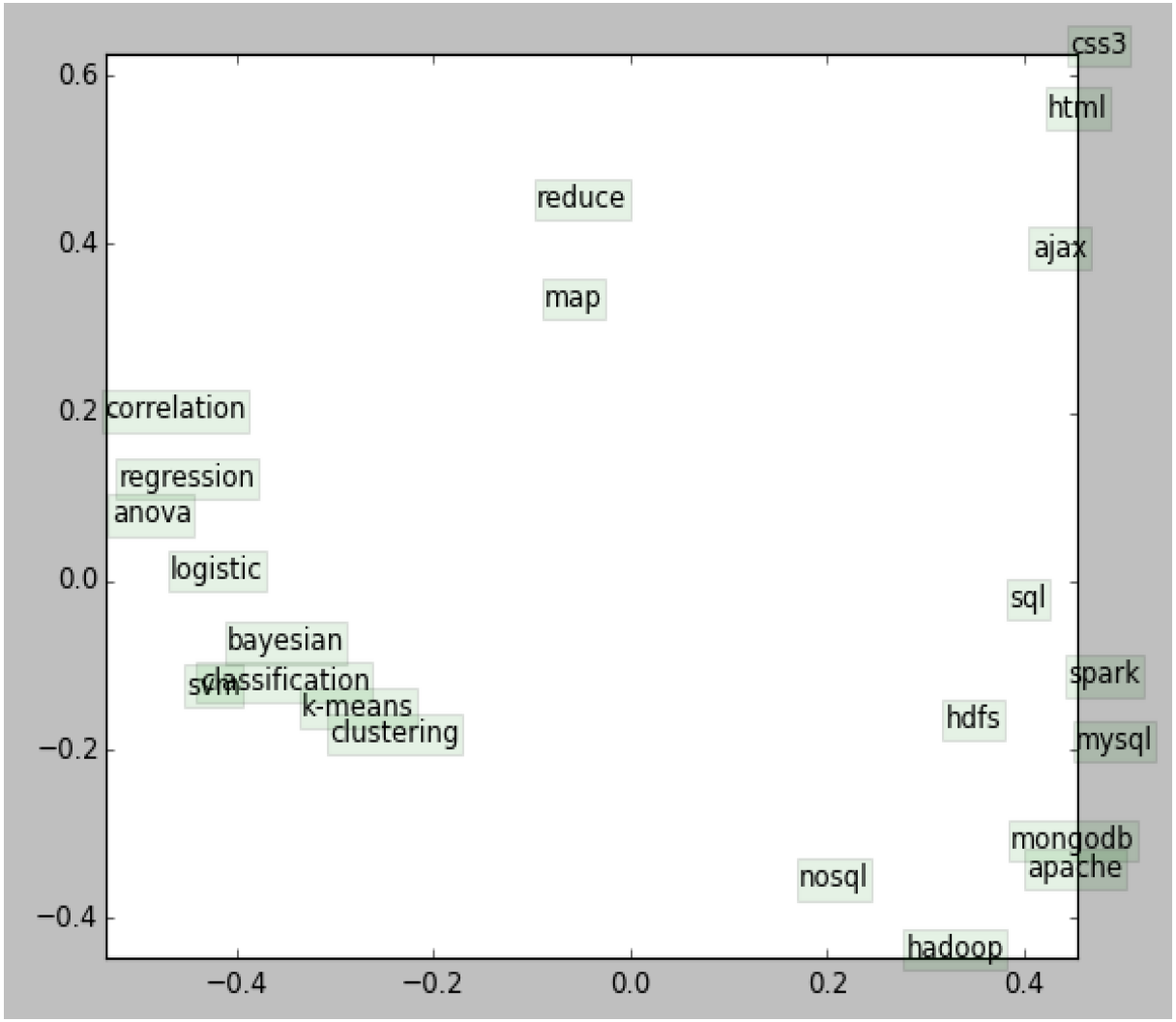
- Syntactically

Vector(apple)- vector(apples) = vector(car)-vector(cars)

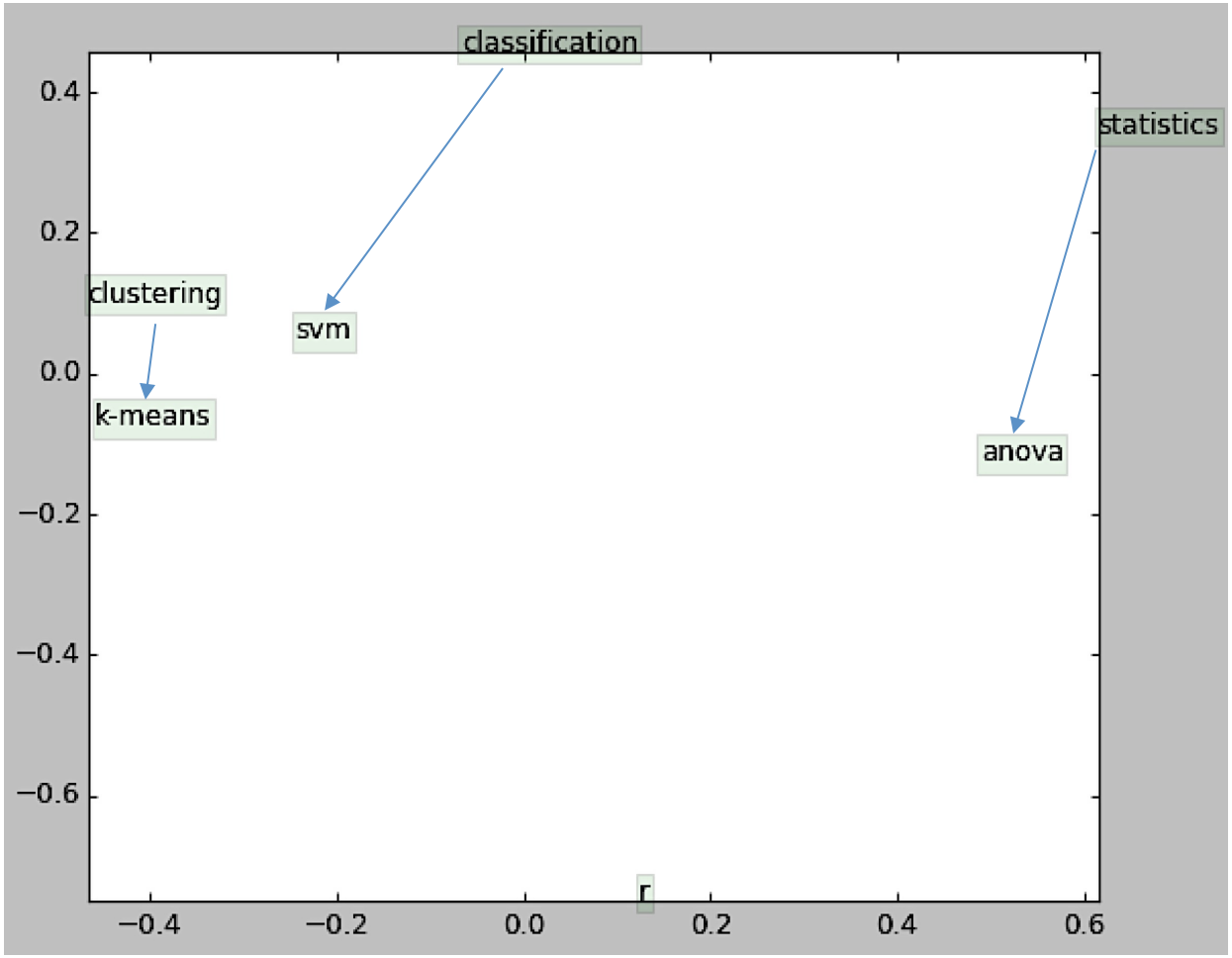
Vec(built) – Vec(build) = Vec(developed) – Vec(develop)

BYOR

Classifying Data Science Keywords



More sophisticated relationships



Let's make the Phrase Vectors!

- https://bitbucket.org/yunazzang/aiwiththebest_byor

```
AERINs-MacBook-Pro:bitbucket_BYOR aerin$ python WTB_phrase_similarity.py
Loading the data file... Please wait...
Successfully loaded 3.6 G bin file!
##################################################################
##################################################################
########## WELCOME TO THE PHRASE SIMILARITY CALCULATOR ##########
##################################################################
##################################################################
Type the phrase1: How Trump Would Stimulate the U.S. Economy
Type the phrase2: Trump unveils plan to revitalize America's economy
##################################################################
Similarity Score:  0.732695
##################################################################
Type the phrase1: Let's make America great again
Type the phrase2: We will make America prosperous and powerful
##################################################################
Similarity Score:  0.670172
##################################################################
Type the phrase1: Let's make America great again
Type the phrase2: We are going to build a beautiful wall
##################################################################
Similarity Score:  0.438079
##################################################################
```