Twitter User Classification

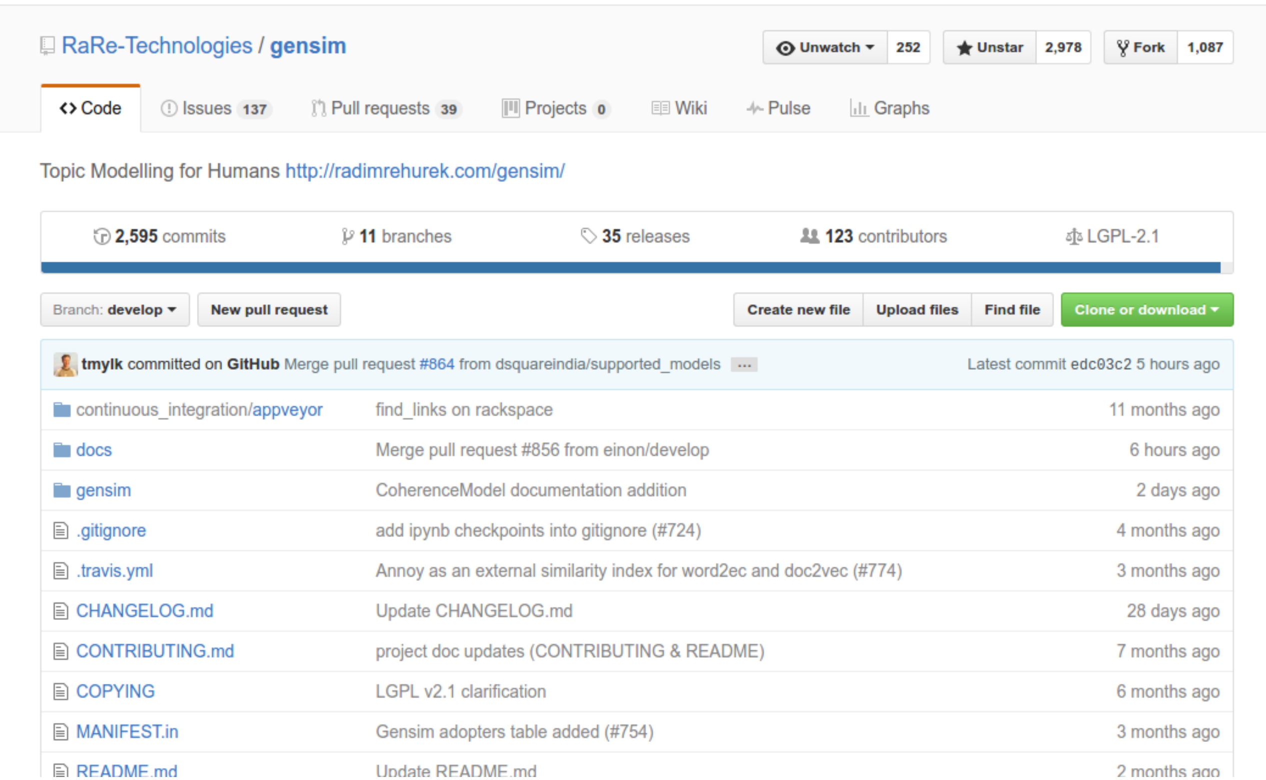
Devashish Deshpande

Undergraduate Student

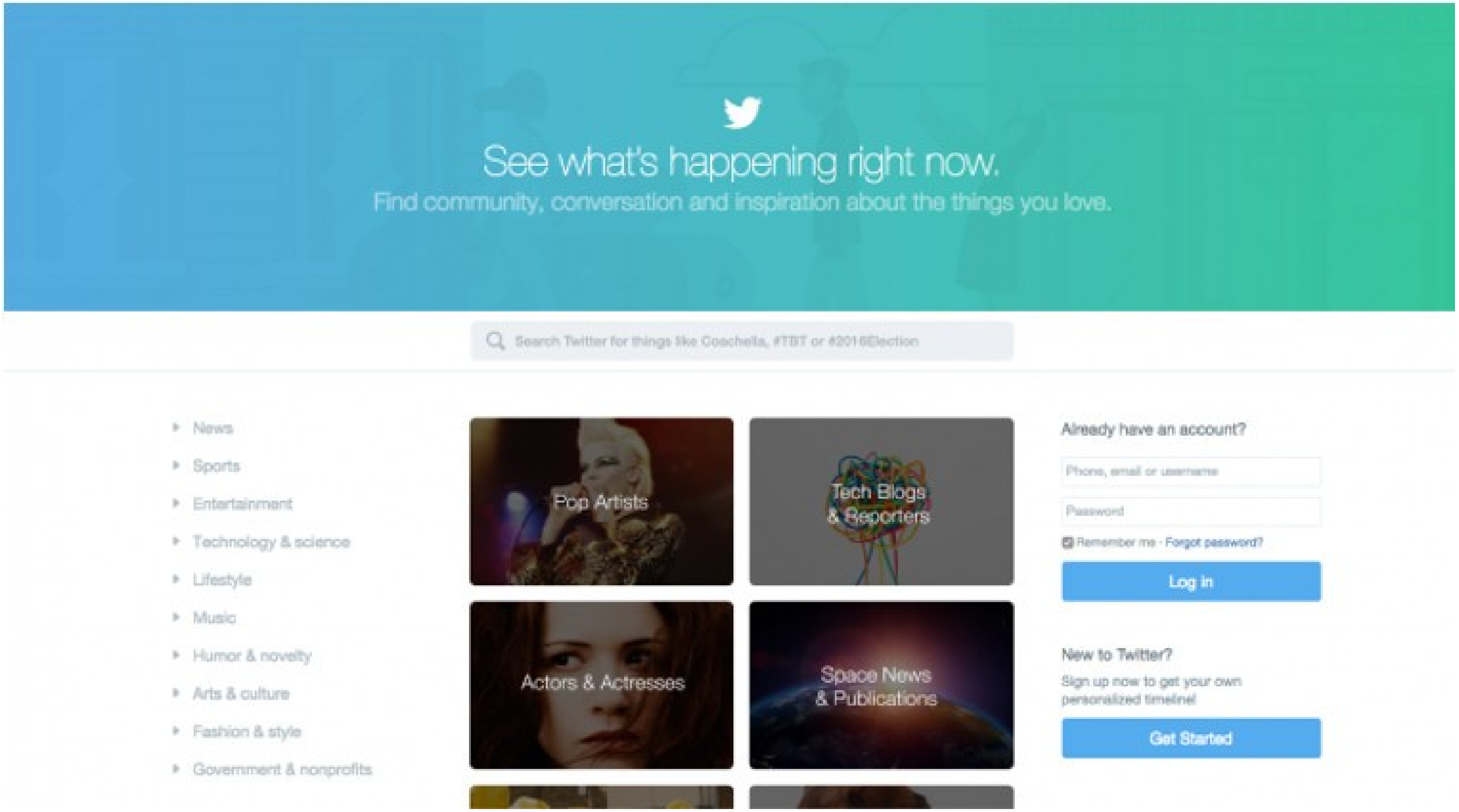
RaRe Technologies Incubator Program

GitHub: dsquareindia
Blogs: https://rare-technologies.com/blog/

Gensim: Topic Modeling in python



Business Problem



- Millions of users on twitter
- Need to make category-wise recommendations
- How classify a user into a category?

Can you classify these users?

Soirée parfaite ! 🇫🇷👑

50 goles en 81 partidos de Liga con el Atleti. El trabajo da sus frutos 💪

Nouvelles couleurs et toujours le même objectif : GOLAZO ⚽ #GriziPums

Can you classify these users?

Soirée parfaite ! 🇫🇷👑

50 goles en 81 partidos de Liga con el Atleti. El trabajo da sus frutos 💪

Nouvelles couleurs et toujours le même objectif : GOLAZO ⚽ #GriziPums

Can you classify these users?

The band have FOUR @mtvema nominations! Voting only takes a couple of clicks (and you can do as often as you like)

In case you missed it - 18 new #AHFODtour shows confirmed for US & Canada in Aug/Sep/Oct 2017

Face à la demande exceptionnelle @Coldplay ajoute une date supplémentaire le 16 juillet 2017 au @StadeFrance

Can you classify these users?

The band have FOUR @mtvema nominations! Voting only takes a couple of clicks (and you can do as often as you like)

In case you missed it - 18 new #AHFODtour shows confirmed for US & Canada in Aug/Sep/Oct 2017

Face à la demande exceptionnelle @Coldplay ajoute une date supplémentaire le 16 juillet 2017 au @StadeFrance

Can you classify these users?

Vous venez à la #PyConFr ? Soyez sympa: pensez à vous inscrire qu'on sache combien on sera: http://2016.pycon.fr/ !#Python #PyCon #Rennes

Regardez ces beaux t-shirts ! https://2016.pycon.fr/t-shirts.html A sérigraphier soi-même pour ceux et celles qui veulent, le jour de l'évènement !

Qui est de PyCon la semaine prochaine à Rennes? Je serais sur place jeudi. Un AFPyro quelque part le jeudi soir? #python #pyconfr

Can you classify these users?

Vous venez à la #PyConFr ? Soyez sympa: pensez à vous inscrire qu'on sache combien on sera: http://2016.pycon.fr/ !#Python #PyCon #Rennes

Regardez ces beaux t-shirts ! https://2016.pycon.fr/t-shirts.html A sérigraphier soi-même pour ceux et celles qui veulent, le jour de l'évènement !

Qui est de PyCon la semaine prochaine à Rennes? Je serais sur place jeudi. Un AFPyro quelque part le jeudi soir? #python #pyconfr

Human mind is a great classifier!

How can we model our thought process?

Creating our dataset

- Created a toy dataset consisting of 8 categories
- Each category consists of:
  - Popular individuals in that category
  - Popular magazine/tv channel handles
- Both tweet in a different style
- Need to capture all kinds of tweets

Creating our dataset

Neil deGrasse Tyson

Not that anybody asked, but on Mercury, I'm 240 years old, and on Saturn, I'm just 2.

If ComicCon people ruled the world, international conflicts would be resolved entirely by plastic light saber fights in bars

NASA

What science experiments did crew members on @Space_Station work on this week? Find out here

We're reducing carbon emissions from aviation by creating aircraft designed to dramatically reduce fuel use & noise: http://go.nasa.gov/2e1BRpn

Both should be classified into "Science"

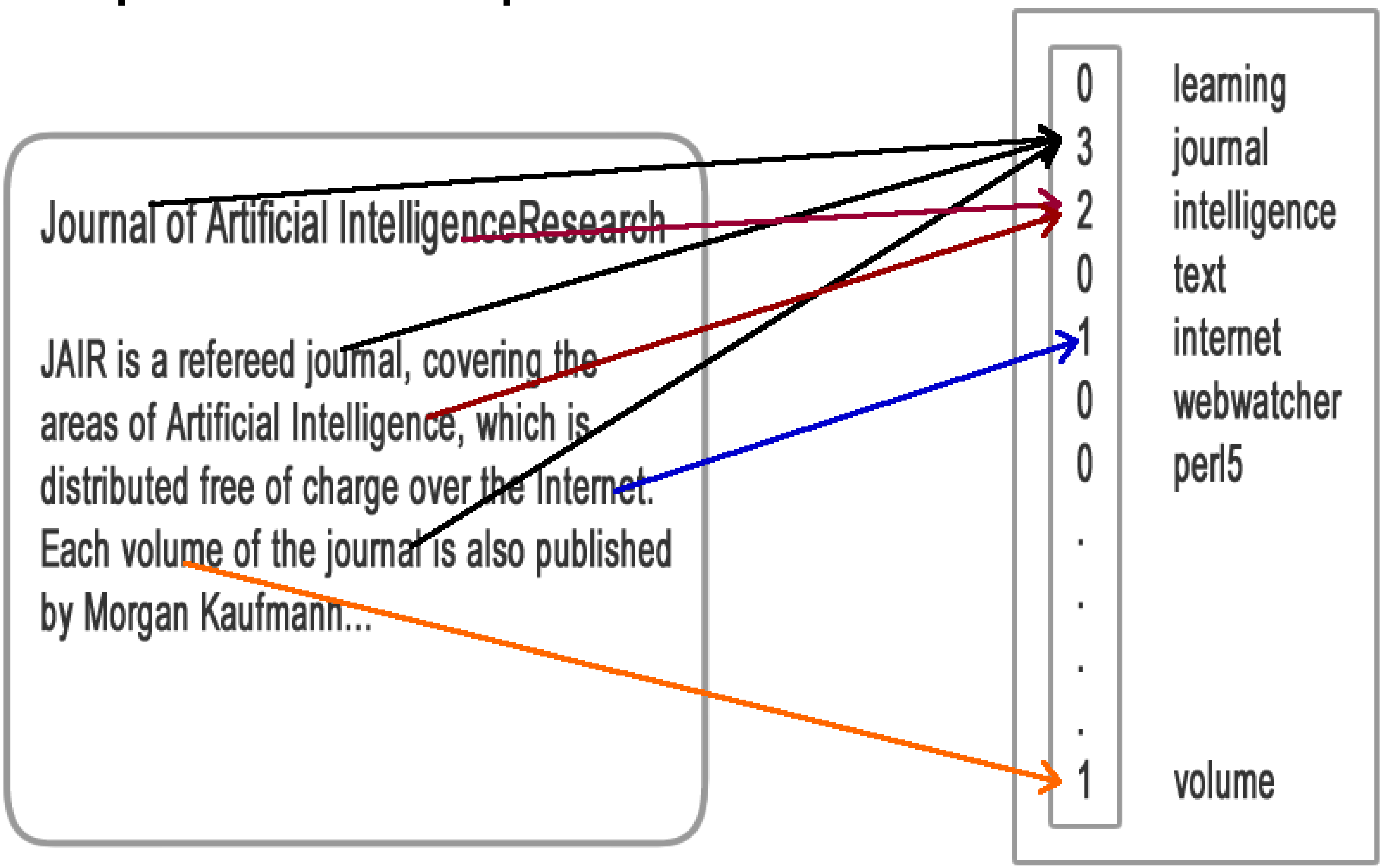
Creating our dataset

- Our categories:
  - Tech
  - Business & CEOs
  - Entertainment
  - Science
  - Fashion, Travel and Lifestyle
  - Sports
  - Music
  - Politics

- Music
  - Iron Maiden
  - David Guetta
  - Rolling Stone magazine
  - MTV Music and more...

Bag of words

- Simple count representation



Bag of words

- Can be great in our case
- Tweets have distinct words which are used often

```
In [49]: most_influential_words(clf_count, count_vectorizer, category_index=7)  # Top words for science category
Out[49]: [u'science',
          u'earth',
          u'scientist',
          u'space',
          u'alien',
          u'study',
          u'dystrophaeus_fossiltime',
          u'bone',
          u'book',
          u'astronaut']
```

Image from accompanying notebook

TF-IDF

- Slightly more complex
- Normalizing using number of docs the term appears in

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Can we try something else?

We are trying to classify documents into topics.

Can we find out the topics using topic modeling and then classify?

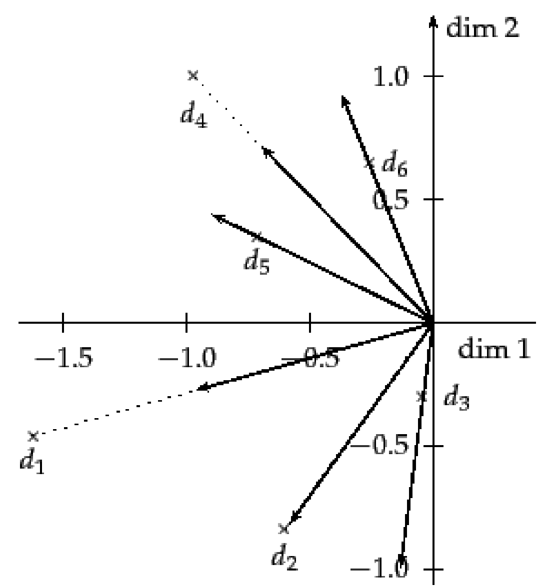
Topic modeling

- Gensim has a number of topic modeling techniques available! We’ll be using
  - Latent Semantic Indexing
  - Hierarchical Dirichlet Process
  - Latent Dirichlet Allocation

Latent Semantic Indexing

- Uses SVD
- Maps terms and documents to a latent semantic space

	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1



Images taken from Stanford IR book

Latent Semantic Indexing

- Can rank topics automatically
- Need to provide num_topics: the number of latent dimensions

```
In [8]: lsi.print_topics(2)

Out[8]: [(0,
          u'0.703*"trees" + 0.538*"graph" + 0.402*"minors" + 0.187*"survey" + 0.061*"system" + 0.060*"ti
        me" + 0.060*"response" + 0.058*"user" + 0.049*"computer" + 0.035*"interface"'),
         (1,
          u'-0.460*"system" + -0.373*"user" + -0.332*"eps" + -0.328*"interface" + -0.320*"response" + -
        0.320*"time" + -0.293*"computer" + -0.280*"human" + -0.171*"survey" + 0.161*"trees"')]
```

Image taken from Topics_and_Transformations notebook trained on deerwester.mm corpus

Hierarchical Dirichlet Process

- Fully unsupervised!
- No need of num_topics parameter
- Determines number of topics through posterior inference
- Non-parametric generalization of LDA
- Can sometimes fail to capture granularity in topics

Latent Dirichlet Allocation

- “Celebrity topic model”. One of the most popular topic modeling algorithms
- Generative process
- Each document is a mixture of topics
- Each topic generates words
- Needs num_topics for fitting

Topic-Word coloring with LDA

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which

Image taken from original LDA paper by Blei et al

What is a good LDA model?

- Come up with interpretable topics
- Infer topic distribution

Football LDA model

Topic 0: mourinho, red_devils, old_trafford, bad_team...

Topic 1: wenger, henry, invincibles.....

Topic 2: aguero, etihad, england, premier_league

Topic 3: blues, football, roman, bridge

We have so many topic models. Great!

But how do we compare them?

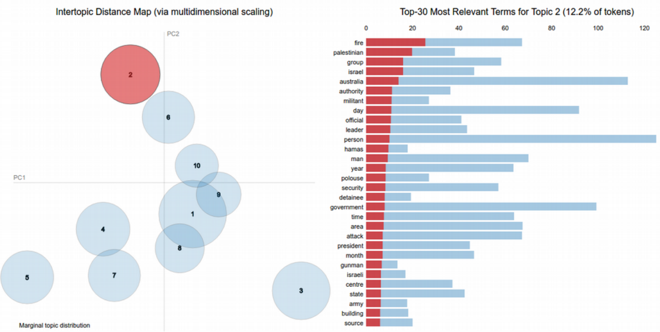
```
In [18]: hdpmodel.show_topics()
Out[18]: [u'topic 0: 0.004*collapse + 0.004*afghanistan + 0.004*troop + 0.003*force + 0.003*government + 0.002*benefit + 0.0
02*operation + 0.002*taliban + 0.002*time + 0.002*today + 0.002*ypre + 0.002*tourism + 0.002*person + 0.002*help +
 0.002*wayne + 0.002*fire + 0.002*peru + 0.002*day + 0.002*united_state + 0.002*hih',
 u'topic 1: 0.003*group + 0.003*government + 0.002*target + 0.002*palestinian + 0.002*end + 0.002*terrorism + 0.002
*cease + 0.002*memorandum + 0.002*radio + 0.002*call + 0.002*official + 0.002*path + 0.002*security + 0.002*wayne +
 0.002*attack + 0.002*human_right + 0.001*four + 0.001*gunman + 0.001*sharon + 0.001*subsidiary',
 u'topic 2: 0.003*rafter + 0.003*double + 0.003*team + 0.002*reality + 0.002*manager + 0.002*cup + 0.002*australia
 + 0.002*abc + 0.002*nomination + 0.002*user + 0.002*freeman + 0.002*herberton + 0.002*lung + 0.002*believe + 0.002
*injury + 0.002*steve_waugh + 0.002*fact + 0.002*statement + 0.002*mouth + 0.002*alejandro',
 u'topic 3: 0.003*india + 0.003*sector + 0.002*anthony + 0.002*interview + 0.002*suicide_bomber + 0.002*union + 0.0
02*marconi + 0.002*imprisonment + 0.002*document + 0.002*mood + 0.002*remember + 0.002*repair + 0.002*vicki + 0.001
*training + 0.001*dressing + 0.001*government + 0.001*indian + 0.001*law + 0.001*convention + 0.001*pair',
```

Manual comparison

- Qualitative analysis
- Can get ugly

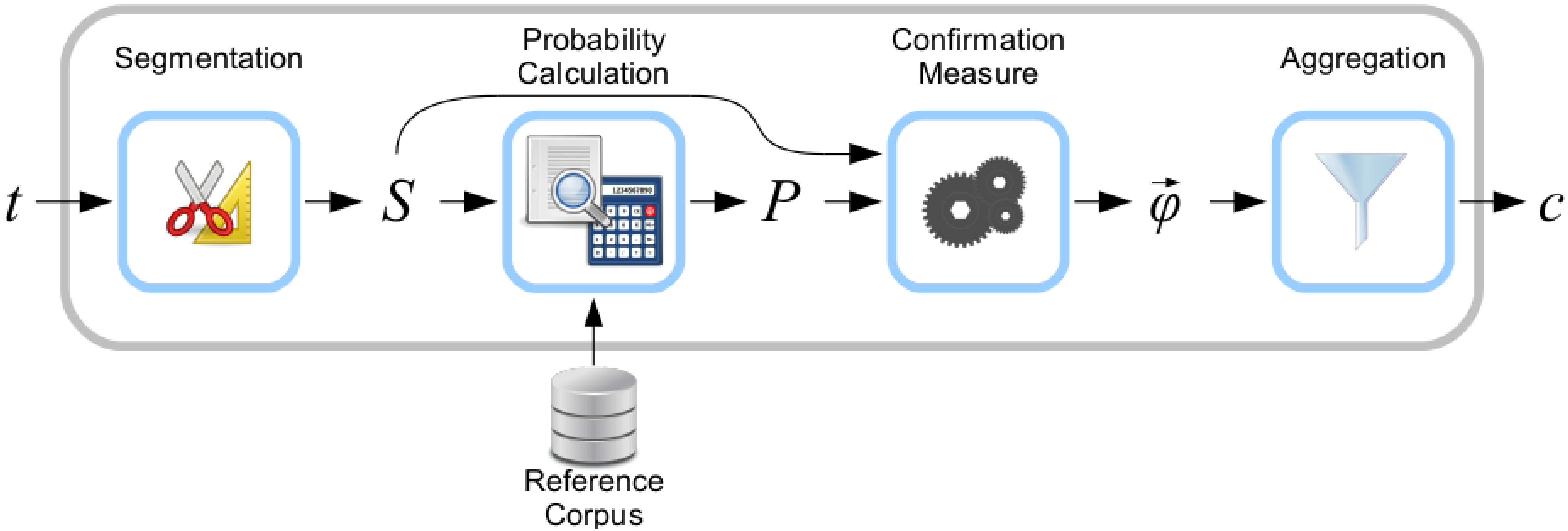
```
 u'0.483*"travel" + 0.235*"time" + 0.229*"thank" + 0.207*"day" + 0.192*"world" + 0.192*"today" + 0.180*"year" + 0.
159*"france" + 0.149*"video" + 0.145*"week"'),
 (1,
 u'0.823*"travel" + -0.149*"today" + -0.147*"thank" + -0.124*"year" + -0.122*"day" + -0.112*"time" + -0.103*"vide
o" + -0.096*"person" + -0.093*"trump" + -0.083*"week"'),
 (2,
 u'-0.875*"france" + -0.264*"paris" + 0.166*"travel" + -0.155*"france_france" + 0.087*"thank" + -0.063*"debate" +
 -0.059*"wine" + -0.058*"ban" + 0.057*"time" + -0.056*"food"'),
 (3,
 u'0.461*"ceo_medium" + 0.454*"number_ceo" + 0.454*"congratulation_number" + 0.420*"congratulation_ceo" + 0.399*"m
edium_socialceos" + 0.124*"ceo" + 0.055*"ceos_medium" + 0.034*"medium_congratulation" + 0.034*"ceo_number" + 0.034
*"company"'),
 (4,
 u'-0.918*"recipe" + 0.141*"trump" + -0.116*"cheese" + -0.083*"gbbo" + -0.076*"chef" + -0.065*"pizza" + -0.063*"re
cipe_chicken" + -0.061*"cake" + -0.055*"article" + -0.053*"chocolate"'),
 (5,
 u'0.535*"trump" + 0.354*"india" + 0.166*"recipe" + 0.151*"hillary_clinton" + 0.133*"president" + 0.131*"market" +
 0.130*"debate" + -0.118*"watch" + 0.116*"china" + -0.116*"thank"'),
 (6,
 u'-0.531*"india" + 0.480*"trump" + 0.259*"video" + -0.181*"market" + -0.164*"world" + 0.144*"debate" + 0.134*"hil
lary_clinton" + 0.116*"album" + -0.110*"china" + -0.106*"rate"'),
 (7,
 u'-0.482*"video" + 0.335*"thank" + -0.292*"chart" + -0.279*"album" + -0.215*"india" + -0.193*"song" + -0.159*"deb
ate" + 0.153*"today" + 0.140*"trump" + -0.121*"feat"'),
 (8,
 u'0.254*"chart" + -0.253*"goal" + 0.246*"apple" + -0.230*"watch" + 0.203*"iphone" + 0.153*"google" + 0.147*"scien
ce" + -0.140*"india" + -0.138*"tonight" + -0.125*"game"'),
 (9,
 u'0.723*"chart" + 0.257*"feat" + -0.215*"video" + -0.198*"apple" + 0.175*"remix" + -0.156*"iphone" + 0.138*"than
k" + -0.125*"google" + 0.110*"chart_hmstack" + -0.104*"debate"')]
```

```
 u'topic 8: 0.003*india + 0.003*album + 0.003*cheese + 0.002*chocolate + 0.002*cake + 0.002*summit + 0.002*watch +
 0.002*song + 0.002*person + 0.002*ice_cream + 0.002*minister + 0.002*music + 0.002*water + 0.002*talk + 0.001*tick
et + 0.001*david_bowie + 0.001*pizza + 0.001*time + 0.001*fan + 0.001*president',
 u'topic 9: 0.016*travel + 0.013*recipe + 0.006*ironmaiden + 0.003*rio + 0.002*thank + 0.002*time + 0.002*india +
 0.002*ironmaiden_spain + 0.002*woman + 0.002*year + 0.002*day + 0.002*tonight + 0.002*ironmaiden_italy + 0.002*tea
m + 0.002*cheese + 0.001*birthday + 0.001*china + 0.001*solo + 0.001*city + 0.001*congratulation',
 u'topic 10: 0.004*watch + 0.004*goal + 0.003*theater + 0.003*india + 0.003*slvaus + 0.003*england + 0.002*pokemon
 + 0.002*australia + 0.002*sri_lanka + 0.002*game + 0.002*trainer + 0.002*team + 0.002*sky_sport + 0.002*man + 0.00
2*match + 0.002*mnf + 0.002*everton + 0.002*pakistan + 0.002*test + 0.002*pokemongo',
 u'topic 11: 0.006*video + 0.004*one + 0.004*today + 0.004*thank + 0.003*apple + 0.003*day + 0.003*time + 0.003*wor
ld + 0.003*euro + 0.002*person + 0.002*bigtheparty + 0.002*party + 0.002*everyone + 0.002*night + 0.002*team + 0.00
1*coloring_book + 0.001*life + 0.001*euro_davidguetta + 0.001*friend + 0.001*tomorrow',
 u'topic 12: 0.011*travel + 0.005*world + 0.004*cnnfood + 0.002*earth + 0.002*mission + 0.002*space + 0.002*way +
 0.002*today + 0.002*city + 0.002*time + 0.002*spacecraft + 0.002*sample_return + 0.002*launch + 0.002*dessert + 0.
002*guide + 0.002*day + 0.002*japan + 0.001*view + 0.001*dish + 0.001*park',
 u'topic 13: 0.004*album + 0.004*song + 0.004*video + 0.002*track + 0.002*share + 0.002*year + 0.002*watch + 0.002*
film + 0.001*review + 0.001*stream + 0.001*week + 0.001*science + 0.001*way + 0.001*life + 0.001*star_trek + 0.001*
recording + 0.001*person + 0.001*band + 0.001*book + 0.001*game',
 u'topic 14: 0.004*goal + 0.003*time + 0.002*today + 0.002*day + 0.002*ageofultron_presstour + 0.002*chelsea + 0.00
1*game + 0.001*dystrophaeus_fossiltime + 0.001*diego_costa + 0.001*arsenal + 0.001*watch + 0.001*way + 0.001*man +
 0.001*liverpool + 0.001*shirt + 0.001*year + 0.001*season + 0.001*fossilfriday + 0.001*presstour + 0.001*point',
 u'topic 15: 0.003*space + 0.003*earth + 0.003*year + 0.002*recipe + 0.002*voyager + 0.002*scientist + 0.002*today
 + 0.002*travel + 0.002*day + 0.002*system + 0.002*pic + 0.002*time + 0.001*water + 0.001*star + 0.001*fossil + 0.0
01*summer + 0.001*edge + 0.001*tip + 0.001*idea + 0.001*tonight',
 u'topic 16: 0.005*video + 0.005*canada + 0.004*album + 0.002*song + 0.002*singer + 0.002*china + 0.002*chine + 0.0
02*today + 0.002*frontman + 0.001*avec + 0.001*discussion + 0.001*shanghai + 0.001*pour + 0.001*metallica + 0.001*t
ime + 0.001*thank + 0.001*tour + 0.001*paralympique + 0.001*music + 0.001*day',
 u'topic 17: 0.004*halloween + 0.003*year + 0.003*video + 0.003*vmas + 0.002*today + 0.002*time + 0.002*vma + 0.002
*fan + 0.002*game + 0.002*performance + 0.002*thank + 0.002*day + 0.002*ticket + 0.002*night + 0.002*tonight + 0.00
2*resetind + 0.002*halloween_destination + 0.002*kanye + 0.001*guy + 0.001*techno',
 u'topic 18: 0.004*thank + 0.003*time + 0.003*show + 0.003*video + 0.003*tonight + 0.003*night + 0.002*year + 0.002
```



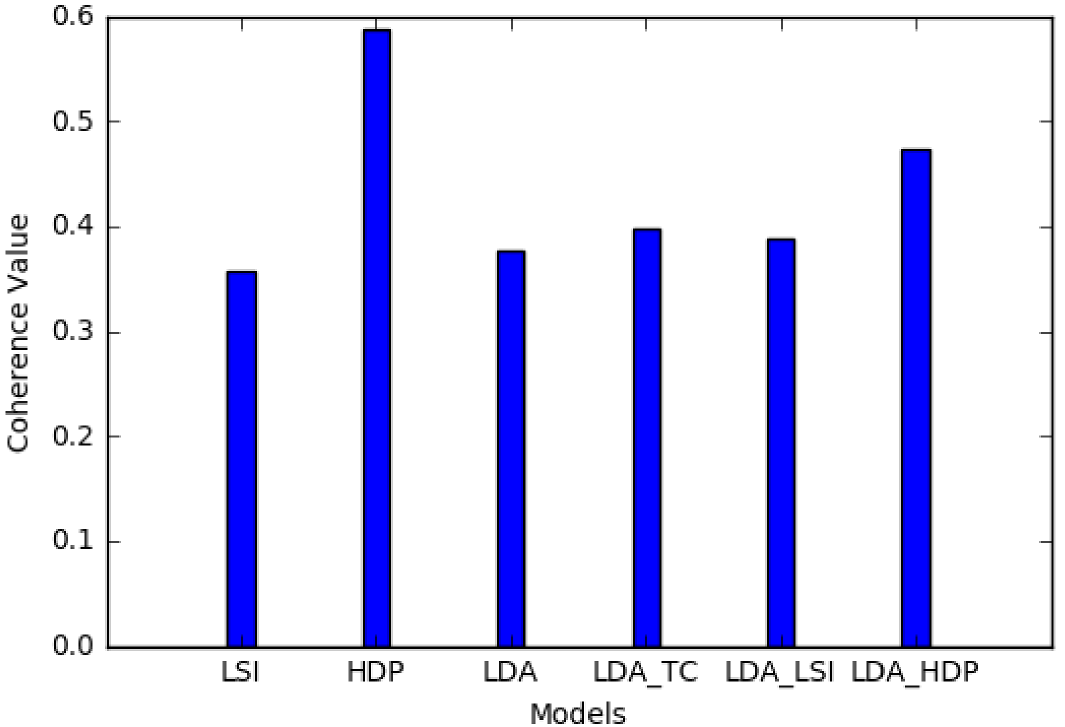
Topic Coherence

- Recently added to gensim
- Based on this paper by Roeder et al



Topic Coherence

- Quantitative
- Enter topics ‘t’ -> Get coherence value ‘c’
- Comparison much easier!



Topic Coherence

- Many other use-cases
  - Selecting num_topics for LDA
  - Rank topics within LDA model (pseudo-LSI)
  - Filtering LDA to improve model

Topic Coherence

- Better LDA -> Better topics -> Better classification

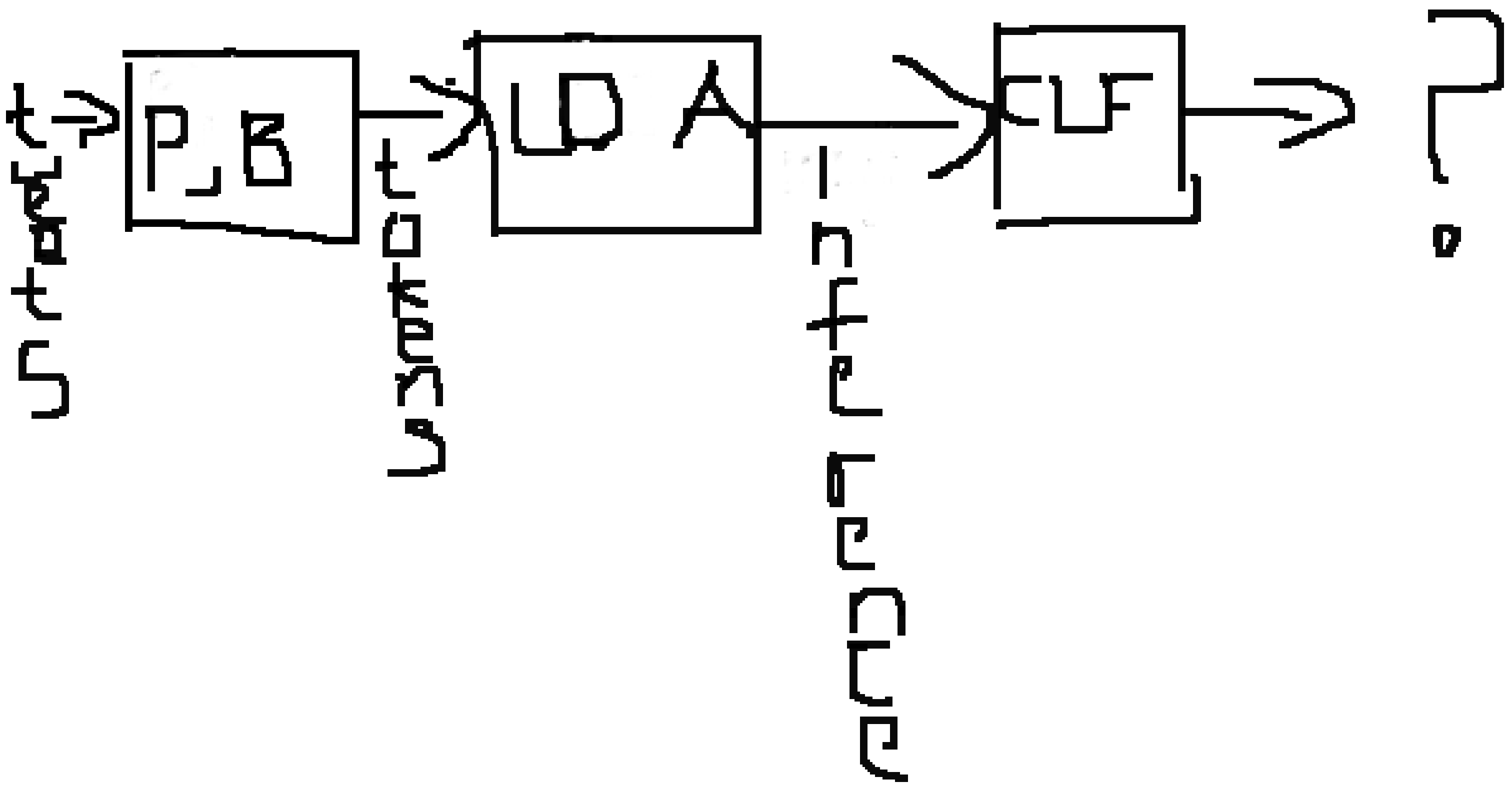
```
[[(u'actor', 0.034688196735986693),
  (u'picture', 0.023163878883499418),
  (u'award', 0.023163878883499418),
  (u'comedy', 0.023163878883499418),
  (u'globe', 0.023163878883499418),
  (u'nomination', 0.023163878883499418),
  (u'actress', 0.023163878883499418),
  (u'film', 0.023163878883499418),
  (u'drama', 0.011639561031012149),
  (u'winner', 0.011639561031012149)],
 [(u'virus', 0.064292949289013482),
  (u'user', 0.048074573973209883),
  (u'computer', 0.040350900997751814),
  (u'company', 0.028173623478117912),
  (u'email', 0.022580226976870982),
  (u'worm', 0.020928236506996975),
  (u'attachment', 0.014534311779706417),
  (u'outlook', 0.012607066546379533),
  (u'software', 0.011909411409069969),
  (u'list', 0.0088116041533348403)],
```

```
[(u'australia', 0.038230610979973961),
 (u'test', 0.030398020440379890),
 (u'day', 0.026478028361575149),
 (u'adam', 0.023237227270639361),
 (u'wicket', 0.018060239149805601),
 (u'match', 0.015652900511647725),
 (u'gilchrist', 0.015206348827236857),
 (u'steve_waugh', 0.014967545716234640),
 (u'south_africa', 0.013902623982144873),
 (u'selector', 0.012332915474867073)],
```

Top topics from topic modeling tutorial on Lee corpus

LDA inference for classification

- We’ll be using our best LDA model’s topic inference for classification

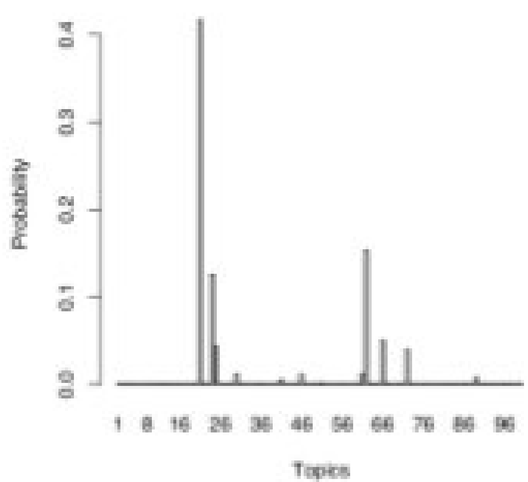


LDA inference for classification

- Using LDA inference to reduce dimensions to num_topics

Real inference with LDA

A 100-topic LDA model was fitted to **17,000 articles from the *Science* journal.**
At right are **the top 15 most frequent words** from the most frequent topics.
At left are the **inferred topic proportions** for the example article from previous slide.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

[D. Blei, **Probabilistic topic models**, *Communications of the ACM*, 2012]

GloVe Vectors

- Algorithm to map words to vectors!

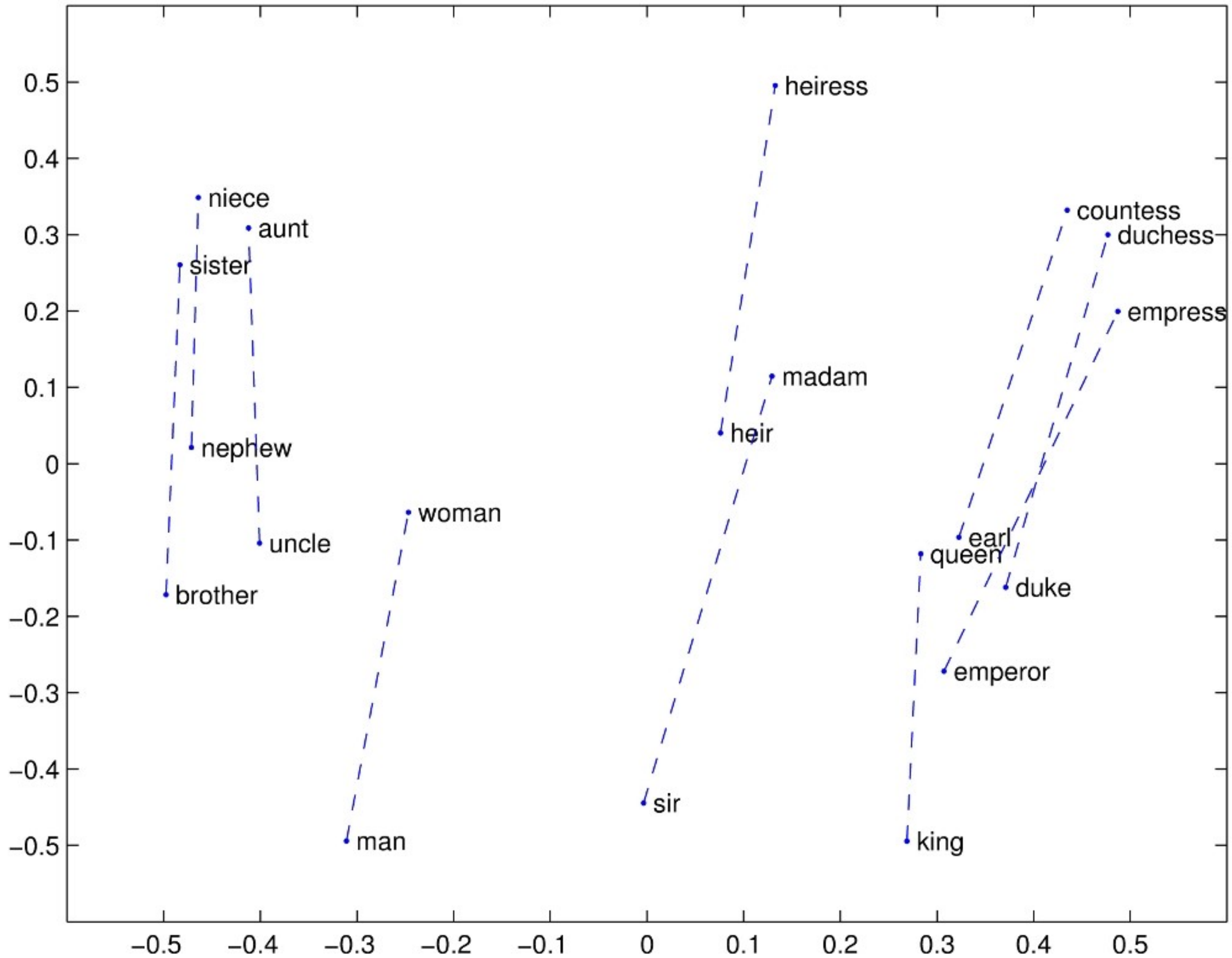
```
In [105]: wv.most_similar(positive=['arsenal'])

Out[105]: [(u'chelsea', 0.9786632061004639),
           (u'liverpool', 0.9681325554847717),
           (u'utd', 0.9317814707756042),
           (u'united', 0.9310426712036133),
           (u'manchester', 0.9226013422012329),
           (u'tottenham', 0.9214313626289368),
           (u'everton', 0.919910728931427),
           (u'barca', 0.9044719934463501),
           (u'swansea', 0.9019159078598022),
           (u'qpr', 0.8937806487083435)]
```

Screenshot from accompanying notebook

GloVe Vectors

- Can do cool stuff!



Picture from glove website

Word2Vec: Makes your heart skip a gram

- How is GloVe different from Word2Vec?

GloVe

- Learns vectors from word-word co-occurences
- “Reconstruction Loss”: Should keep variance while reducing dimensions

Word2Vec

- “Deep learning” predictive model
- CBOW, skip-gram
- Predicting target word given context

Word2Vec vs GloVe

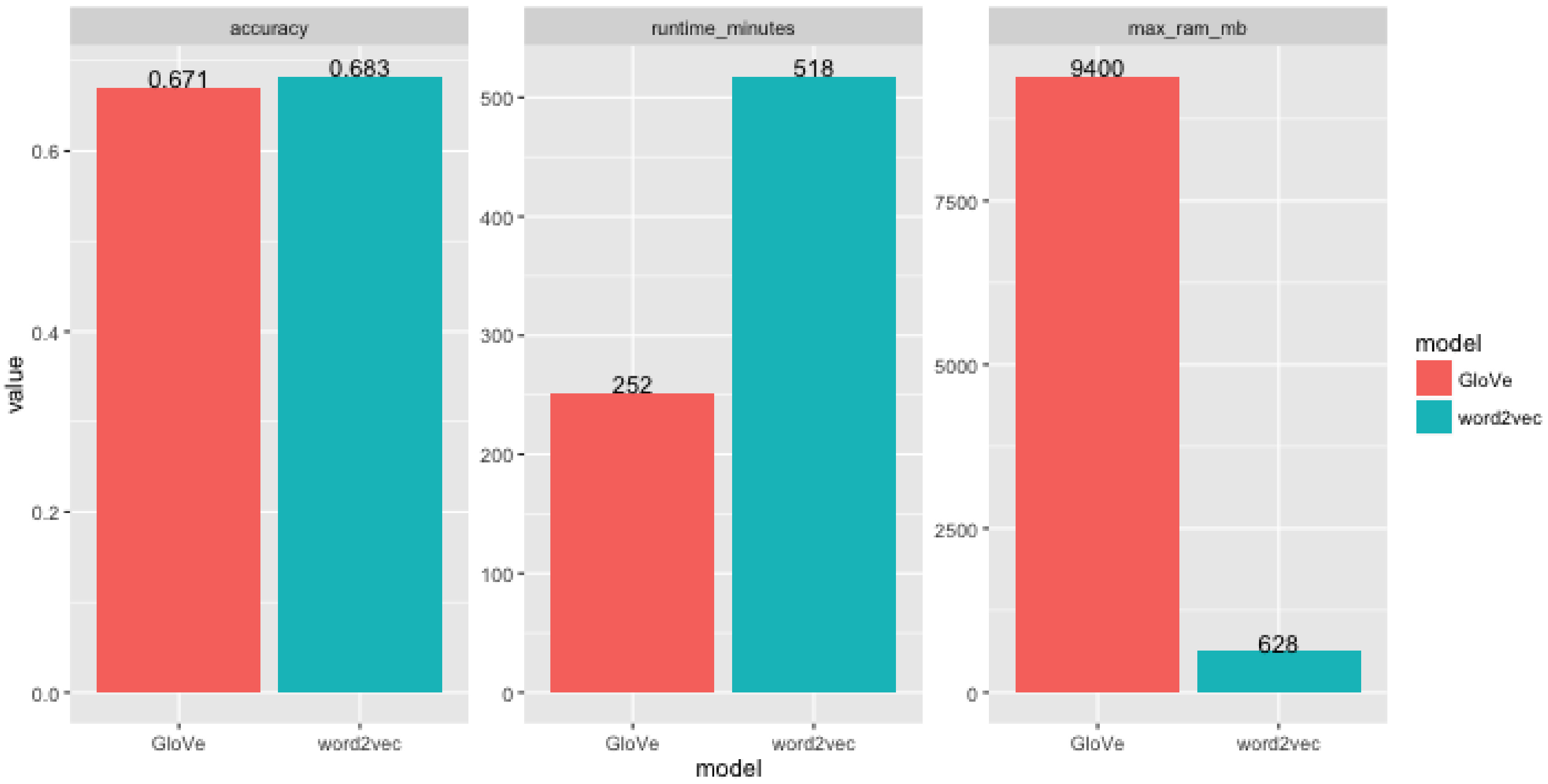


Image from http://dsnotes.com/articles/glove-enwiki

Using GloVe

- Our training time is zero!
- Using pre-trained twitter GloVe vectors by Stanford
- Converting to word2vec format for gensim compatibility using glove2word2vec script

Let’s move on to the tutorial!