TOPIC MODELING FOR TWITTER ACCOUNTS

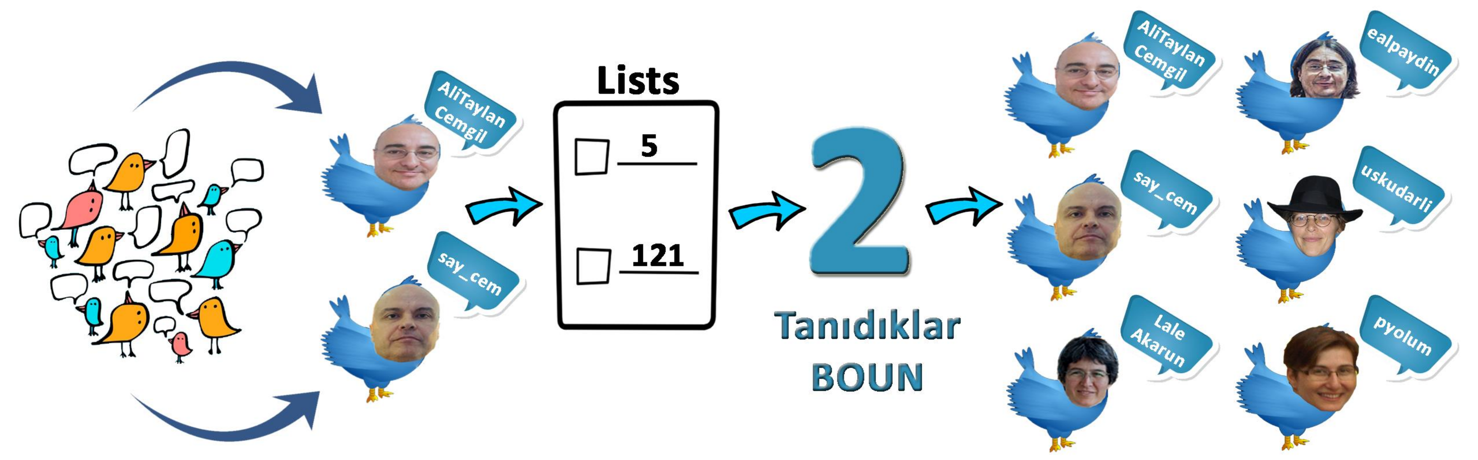
Burak Suyunu Mehmet Akif Çördük

Advisor: Assoc. Prof. Ali Taylan Cemgil

WHAT ARE THEY TWEETING ABOUT

- Makers, scientists, influencers and many other people share their ideas, products and innovations via the most intellectual social network **Twitter**.
- It is hard to find the information about a **topic** in the giant network of Twitter.
- Our aim is to find users who are tweeting about the same **topic**. With this aim we want to bring people interested in the same **community** together.
- In this project, we focused on **maker** communities and **influencers** in the context of computer science, such as **ML**, **Robotics**, **3D Printing**, **Arduino**.
- We worked on **1.118 users** and approximately **3.250.000 tweets**.

DATASET - SIMILAR-TWITTER

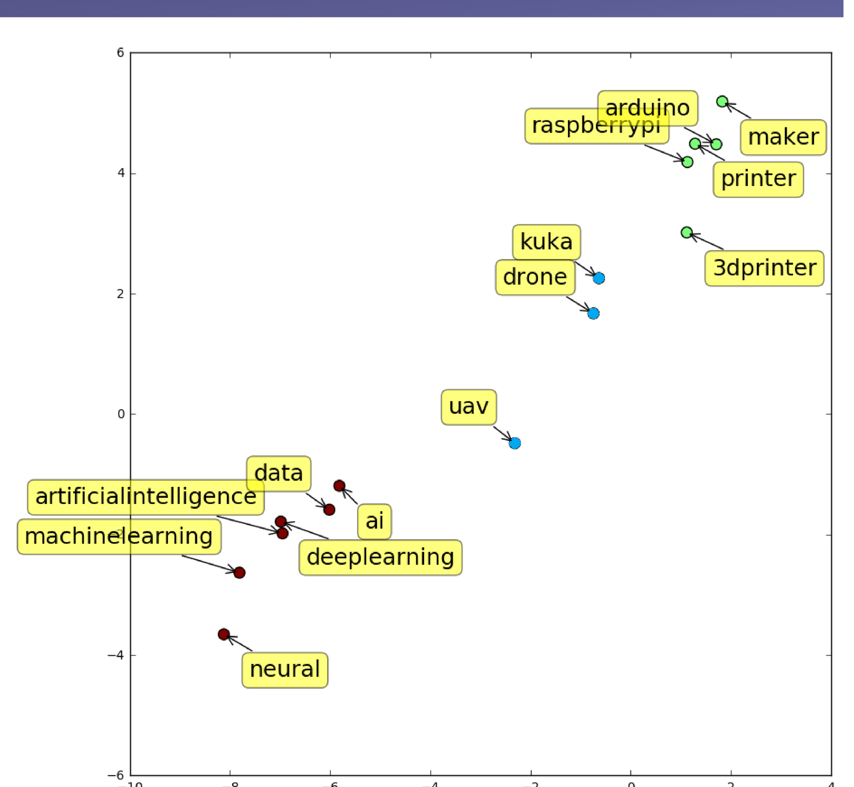


MAINTAINING TWEETS – NLP

- ***Imagination is more important than knowledge: https://Einstein.co #Einstein***
- <u>Remove URLs</u>
  - Imagination is more important than knowledge: #Einstein
- <u>Tokenization</u>
- <u>Stop Words</u>
  - ['imagination', 'important', 'knowledge', 'einstein']
- <u>Remove non-English accounts</u>
- <u>Stemming</u>
  - ['imagin', 'import', 'knowledg', 'einstein']
- <u>Remove words that appears at most 10 times in the whole corpus</u>

CLUSTERING WORDS - WORD2VEC

- **Word2Vec** uses word embedding to map words to a **vector** of real numbers.
- We applied **k-means clustering** to the vectors to see the relevant words together.
- We chose the word at the **center** of the cluster to represent the other words from the same cluster in the word corpus.
- We **normalized** the number of occurrences in the corpus to handle the problem of less frequent words being more important.



TOPIC MODELING

- In **machine learning** and **natural language processing**, a topic model is a type of statistical model for discovering the topics that occur in a collection of documents.
- We know that a document is about a particular topic, we expect particular words to appear more often than others since some words are more related to the subject.
- So we are trying to learn **topic distribution over the vocabulary** or **word distributions of the topics**.

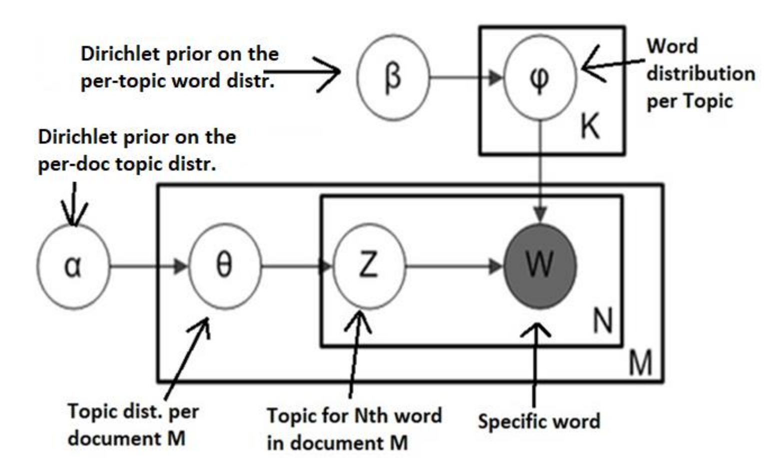
- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Hamsters and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

- **Sentences 1 and 2**: 100% Topic A
- **Sentences 3 and 4**: 100% Topic B
- **Sentence 5**: 60% Topic A, 40% Topic B
- **Topic A**: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … **(Food)**
- **Topic B**: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … **(cute animals)**

LDA (LATENT DIRICHLET ALLOC)

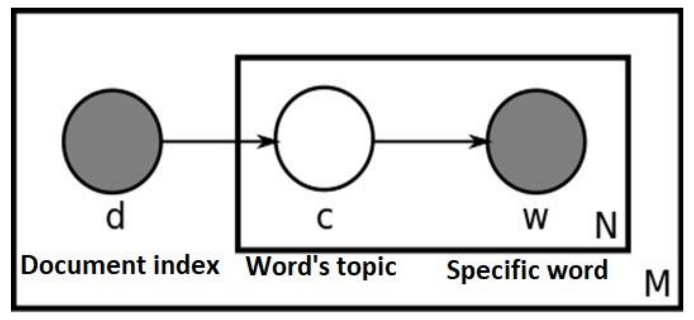
- Assign each word in a document to one of **K topics randomly**
- To obtain a correct distribution, iterate over each document D and for each document iterate over each word W.
- Then, for each topic T reassign the word W to a new topic T':

$$P(Word\ W \mid Topic\ T) * P(Topic\ T \mid Document\ D)$$



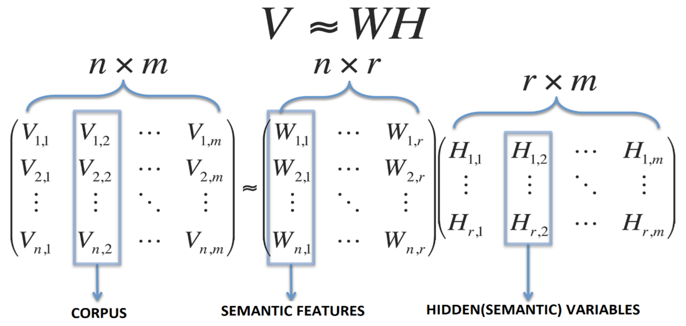
NMF (NON-NEGATIVE MATRIX FACT)

- NMF decomposes the data into two **low rank matrices (W, H)** whose product constitutes the data matrix.
- At each iteration, update W and H with additive update rules to minimize the **squared error** to reach a good decomposition.



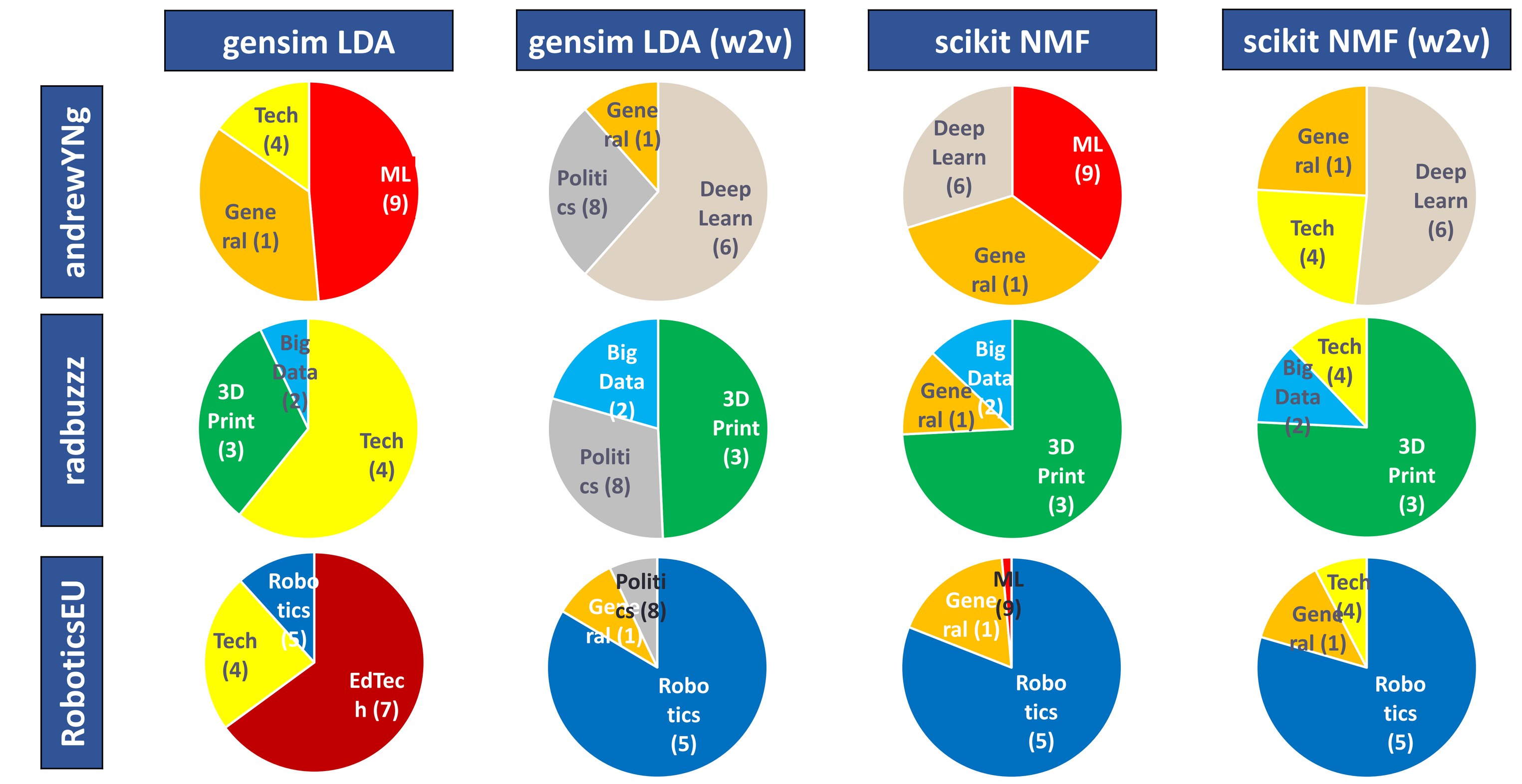
- **NMF** + Kullback-Leibler Divergence + Drichlet priors on distributions => **LDA**
- **NMF** trains much faster than **LDA**

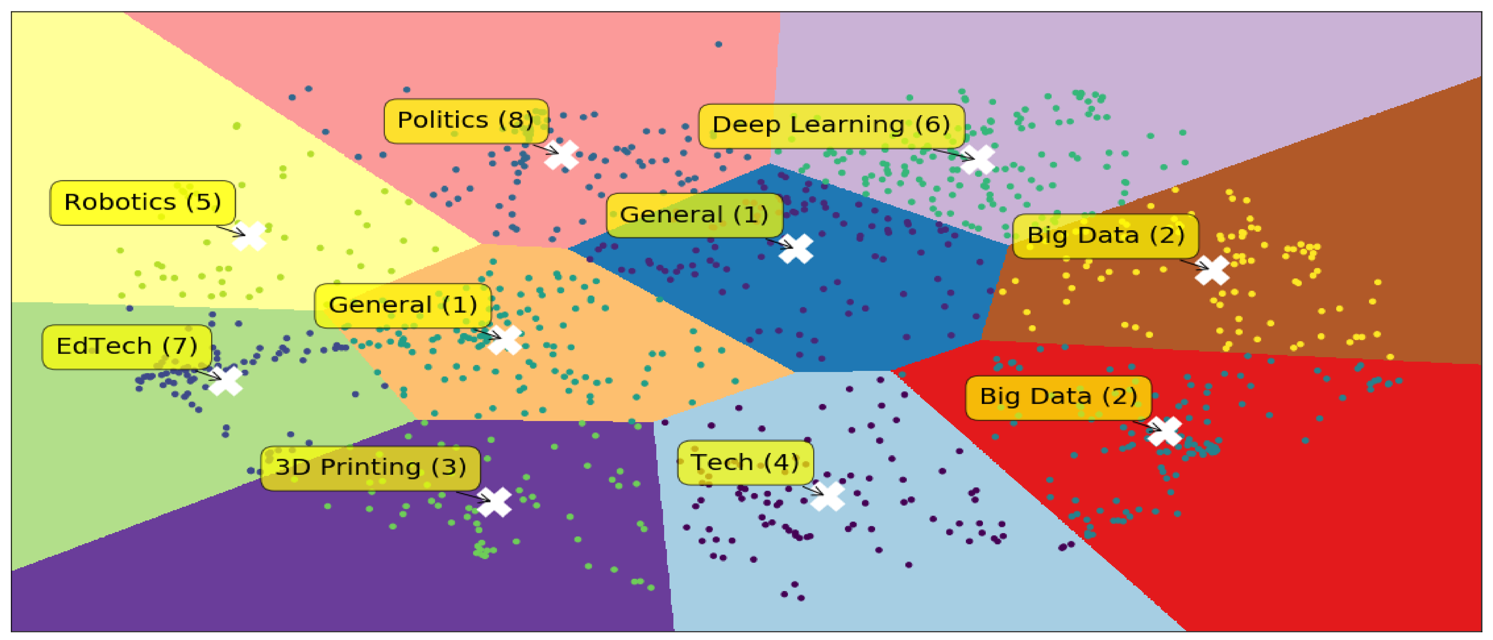
$$V \approx WH$$



RESULTS

	Daily (1)	Big Data (2)	3D Print (3)	Tech (4)	Robotics (5)	Deep Learn (6)	EdTech (7)	Politics (8)	ML (9)	Arduino (10)
gensim LDA	think work time	data bigdata analytics	3D print 3Dprint	tech innov wear	robot manufac automat		stem robot 3dprint	us trump world	datasci data ML	drone arduino robot
scikit LDA	work time think	data bigdata ai	3Dprint 3D print	startup business market	robot manufac us		stem code learn			arduino maker project
scikit NMF	work time look	bigdata analytics data	3Dprint 3D print		robot drone kuka	learn deep neural	edtech stem edchat		datasci ML DeepL	pi rasberrypi raspberry
gensim LDA (w2v)	love day today	bigdata data ai	3Dprint 3D printer	market business startup	robot manufac engineer	learn deep machine	stem code learn	trump year us	datasci data ML	arduino robot project
scikit LDA (w2v)	love day us	data bigdata analytics	3Dprint 3D printer	innov join learn	robot ai techn	learn deep machine	code stem learn	trump us science	datasci data ML	arduino project kit
scikit NMF (w2v)	time day today	bigdata analytics data	3Dprint 3D printer	startup bussiness innov	robot kuka automat	learn deep neural	stem science women	trump vote obama	datasci ML bigdata	arduino kit rasp-pi





CONCLUSION

- The hardest part of our project is the **evaluation** of results. Because all the results we got from topic modeling algorithms needs **human interpretation**. So, to make those interpretation clear and understandable we came up with the idea **of color coded charts**. Even it is hard to interpret, we got very promising and comparable results. While **NMF** generally gives **better** results than **LDA**; **Word2Vec** improved both methods significantly in capturing the general idea.
- All in all, one can find different datasets with **Similar-Twitter** and analyze them with our **Topic Modeling** approaches to create communities.