Bursting the bubble: tweaking Twitter newsfeeds

ALEXANDER OBEID, DENNIS NUNEZ-FERNANDEZ, HÉLÈNE HERMAN

Fighting bias & advocating for empathy

The social media “filter bubble,” is an echo chamber, a ”personal, unique universe of information that you live in online. And what’s in your filter bubble depends on who you are, and it depends on what you do. But the thing is that you don’t decide what gets in. **And more importantly, you don’t actually see what gets edited out.**”

Eli Pariser, Internet activist

Creating a new Twitter newsfeed

Our approach

#1: EMBEDDING OF SENTENCES TO HAVE COORDINATES FOR TWEETS

#2: CREATING A FIRST "NATURAL" RECOMMENDATION ALGORITHM

#3: CREATING A SECOND RECOMMENDATION ALGORITHM TO BURST THE BUBBLE

Our data: Twitter 2012 US Presidential Election

* Over 170,000,000 tweets collected during 3 months leading up to the 2012 US presidential election

* Over 70 themes were identified, among which: Abortion, Afghanistan, BirthCertificate, Hurricane, Iran, Israel, MichelleObama, MiddleClass, MidWest, MarriageEquality, Nationaldebt, Obamacare...

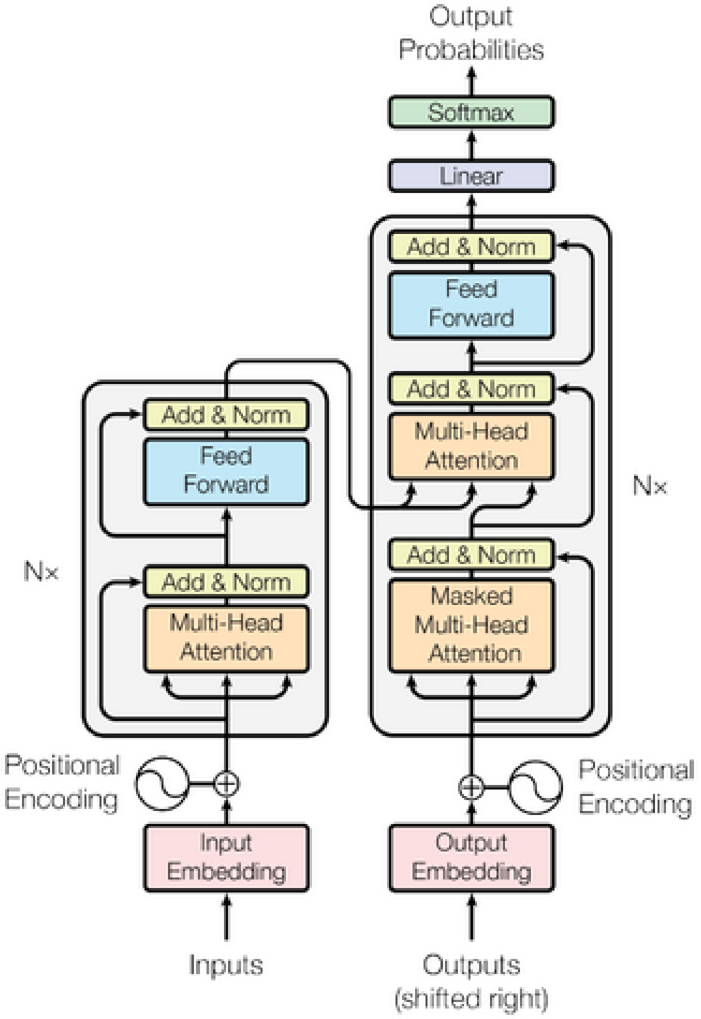
* Dataset containing the information needed for our project, namely: text, description, number of followers, retweet count, language

Sampling & preprocessing

* Out of a sample of 500,000 tweets, we have only kept those where the user had at least 2 tweets to have something to base our recommendation on, ended up with ~145,000 tweets

* We have removed stopwords, tokenized and lemmatized

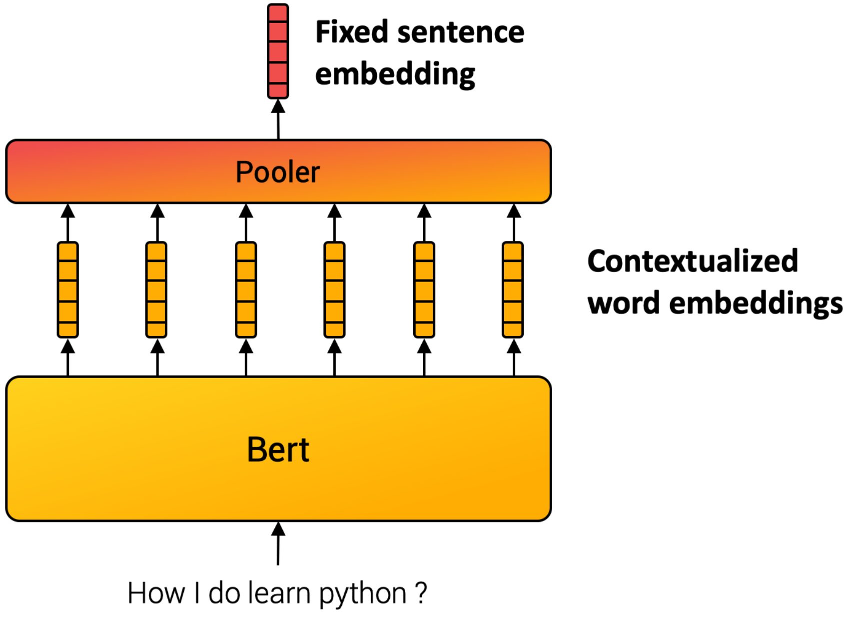
Sentence Embeddings: **SentenceBERT**



BERT (Bidirectional Encoder Representations from Transformers)

SBERT adds a pooling operation to the output of BERT to derive a fixed sized sentence embedding.

The Stanford Natural Language Inference (SNLI) Corpus



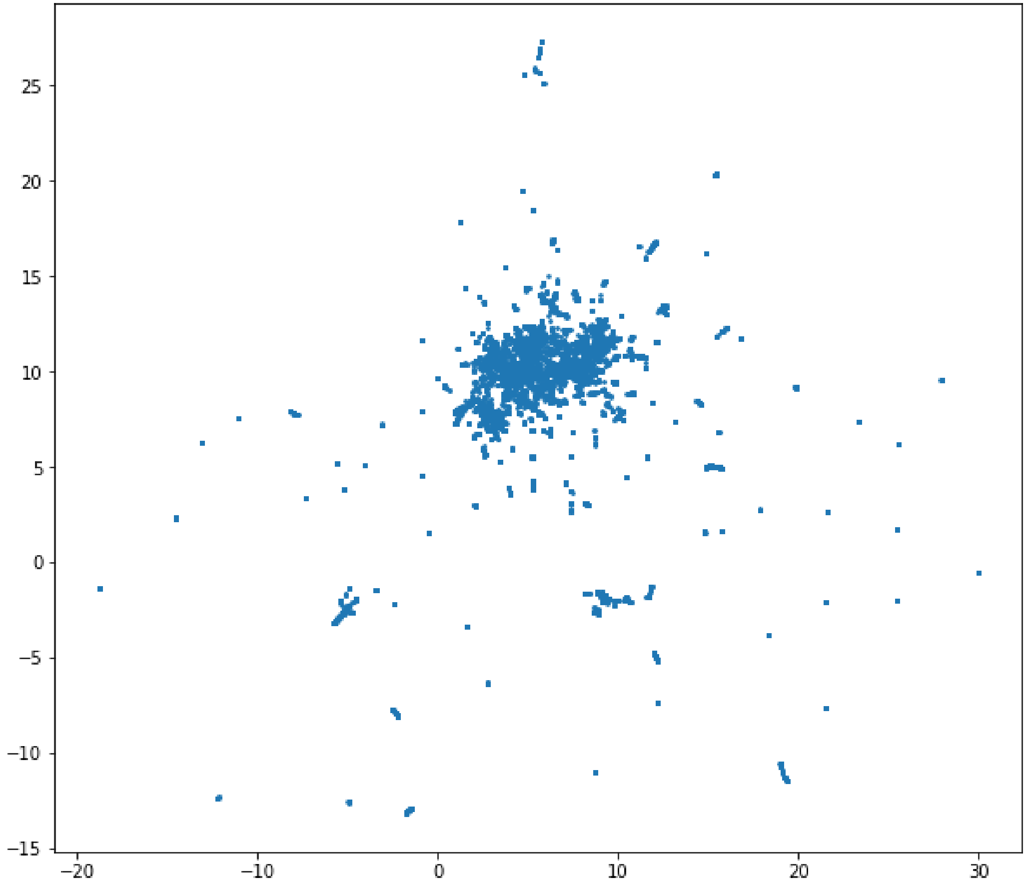
Sentence Embeddings: **SentenceBERT**

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

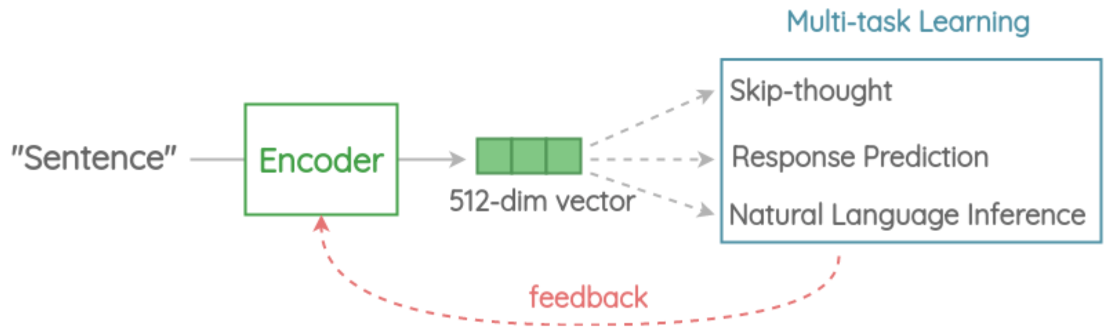
UMAP is a new technique that offers a number of advantages over t-SNE,

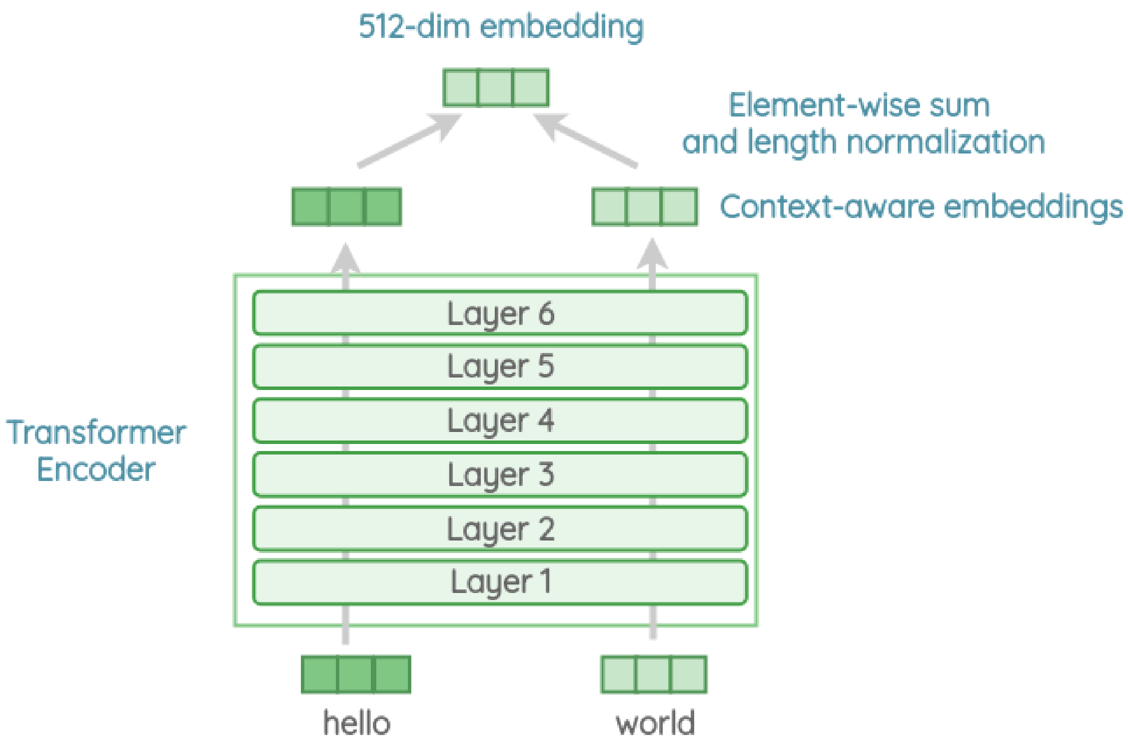
UMAP most notably increased **speed** and better preservation of the data's **global structure**

UMAP much more clearly **separates** these groups of similar categories from each other..



Sentence Embeddings: **Universal Sentence Encoder**



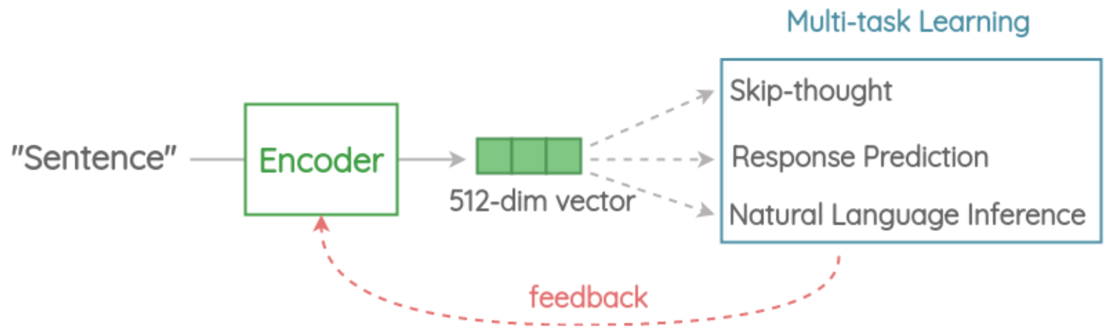


Transformer Encoder

The Stanford Natural Language Inference (SNLI) Corpus

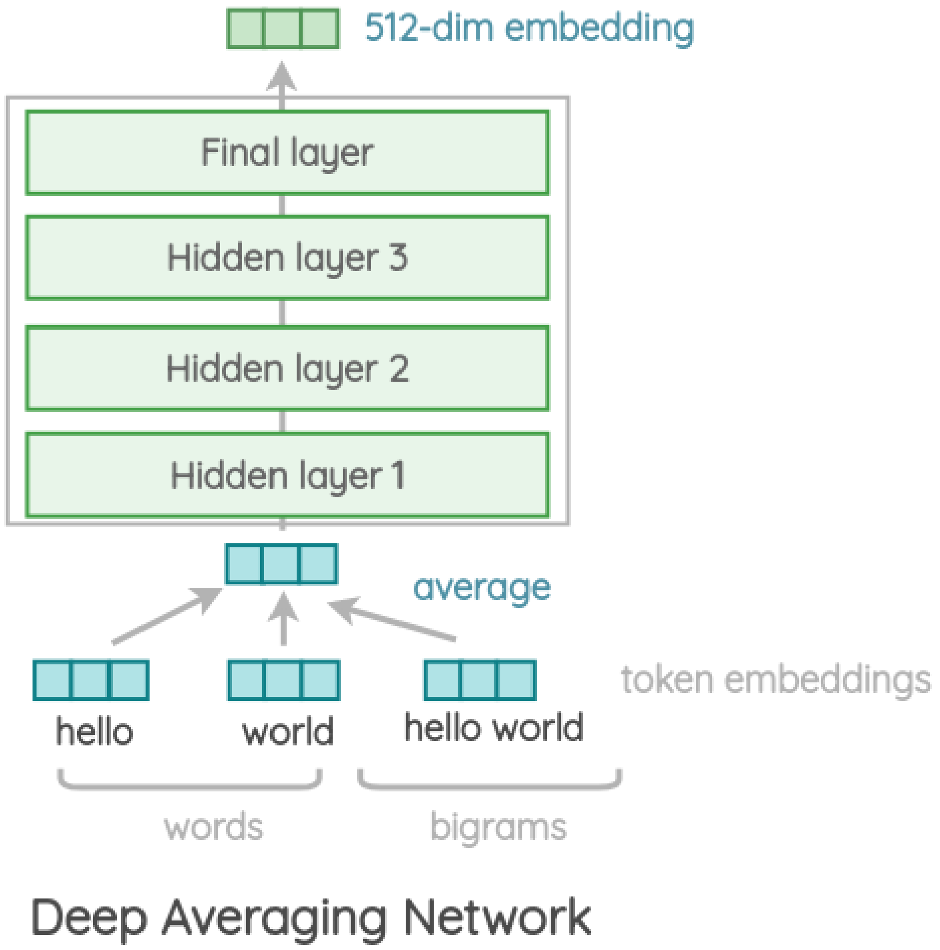
The SNLI corpus is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE)

Sentence Embeddings: **Universal Sentence Encoder**



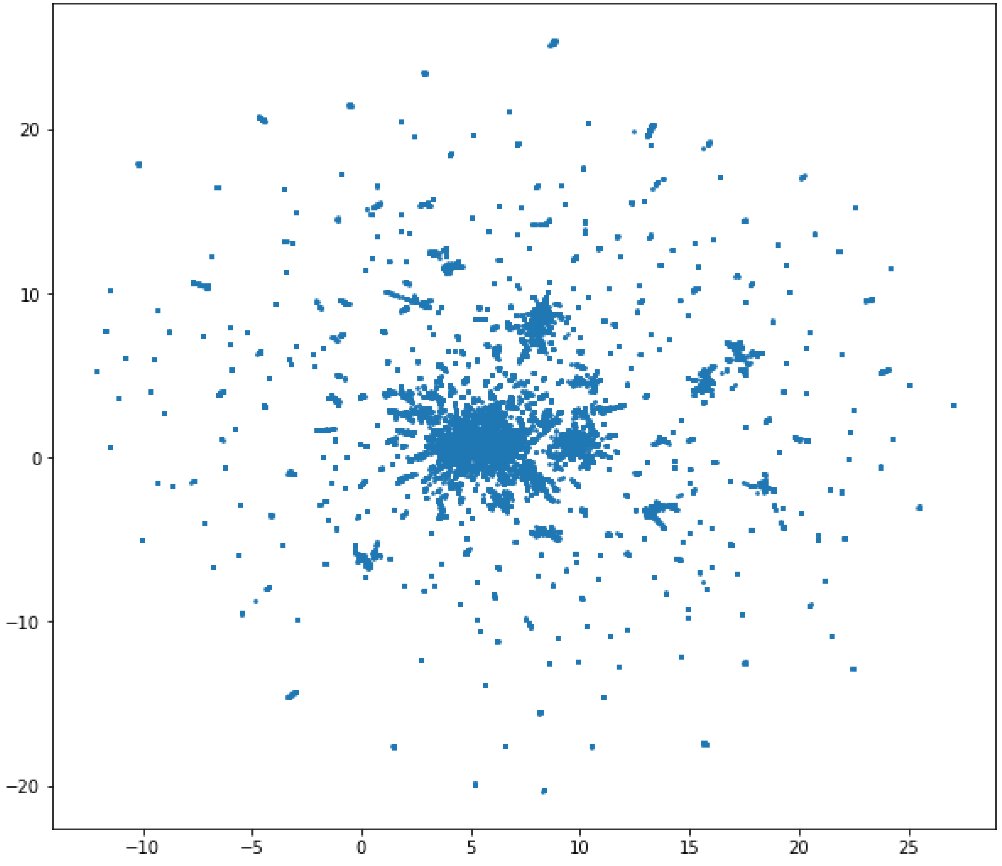
The Stanford Natural Language Inference (SNLI) Corpus

The SNLI corpus is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE)



Sentence Embeddings: **Universal Sentence Encoder**

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction



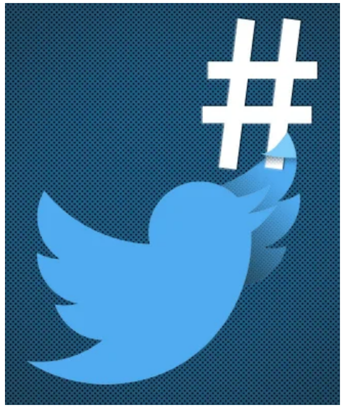
Creating the “Natural” recommendation

Sentence Embedding
(Universal Sentence Encoder)

Similarity Matrix
(Cosine similarity)

Recommend Top N

	1	2	3	4	5	...	N
1	1.000						
2	-0.861	1.000					
3	0.067	-0.080	1.000				
4	0.289	-0.190	0.672	1.000			
5	0.858	0.255	-0.107	0.387	1.000		
...	-0.545	0.990	-0.397	-0.837	0.789	1.000	
N	-0.605	-0.859	0.883	0.791	0.152	0.037	1.000



“Bursting the bubble” recommendation

Sentence Embedding

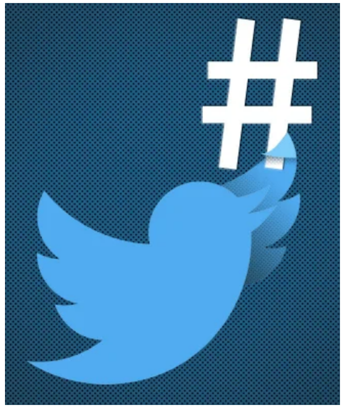
(Universal Sentence Encoder)

Similarity Matrix

(Cosine similarity)

Recommend Last N

	1	2	3	4	5	...	N
1	1.000						
2	-0.861	1.000					
3	0.067	-0.080	1.000				
4	0.289	-0.190	0.672	1.000			
5	0.858	0.255	-0.107	0.387	1.000		
...	-0.545	0.990	-0.397	-0.837	0.789	1.000	
N	-0.605	-0.859	0.883	0.791	0.152	0.037	1.000



"Bursting the bubble" recommendation

Sentence Embedding

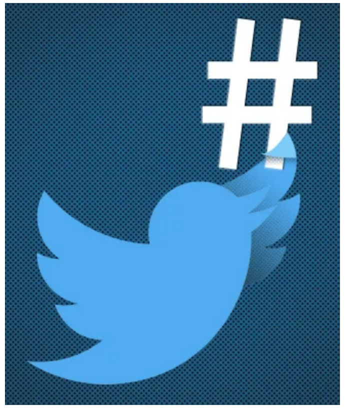
(Universal Sentence Encoder)

Similarity Matrix

(Cosine similarity)

Recommend Last N

	1	2	3	4	5	...	N
1	1.000						
2	-0.8[illegible]	1.000					
3	[illegible].067	-0.080	1.[illegible]				
4	0.289	-0.190	0.672	1.000			
5	0.858	0.255	-0.107	0.387	1.000		
...	-0.545	0.990	-0.397	-0.837	0.789	1.000	
N	-0.605	-0.859	0.883	0.791	0.152	0.037	1.000



“Bursting the bubble” recommendation

Sentence Embedding

(Universal Sentence Encoder)

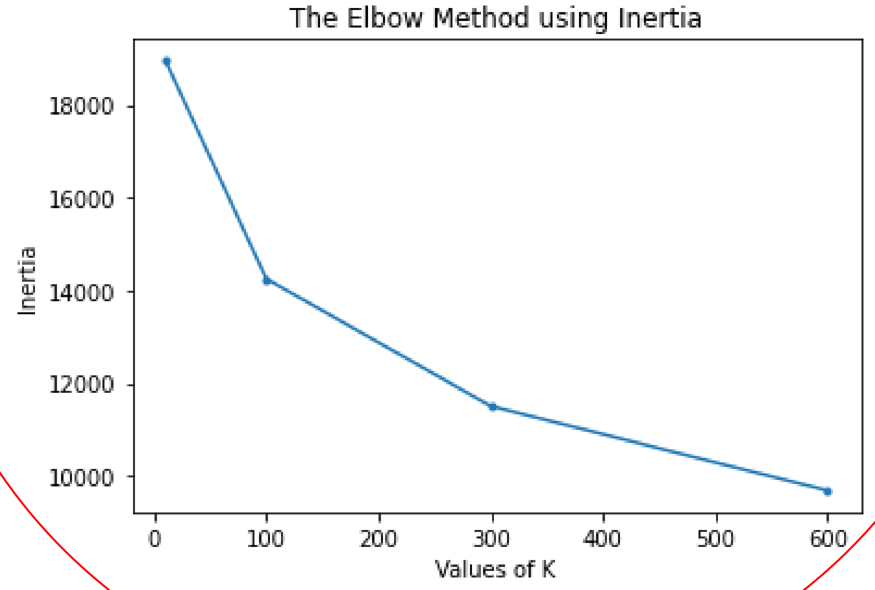
Clustering

(Kmeans)

Similarity Matrix

Recommend

Last N



Example

Starting Tweet

political issues poll vote general election commenter uk

Natural recommendations

political polls archive list political polls general election website

reasons vote united kingdom independence party ukip general election united kingdom

presidential election quiz candidate isidewith

Bursting the bubble

(without clustering) -> rogers black leather sofa guide rogers black leather sofa check price global furniture usa customer

(with clustering) -> respect scarlett election study mexicos

Conclusions

- Evaluating the "bursting" recommendation is really difficult
- Creating a recommendation system for Twitter is possible using sentence embedding and similarity matrices
- To include recommendations that "burst" the bubble of Twitter natural recommendations an extra clustering step is needed
- Maybe other techniques such as n-gramming may be used to "burst" the bubble without using the clustering step
- For further improvement, performance metrics and higher computer capacity are needed, we could also refine recommendations by adding criteria
- The clustering strategy for the "bursting" recommendation seems to be working but needs further tuning on the number of clusters

  Instead of recommending for each tweet, we could recommend to a user based on all its tweets

THANK YOU

DIETER RAMS