TOPIC CLASSIFICATION OF TWEETS

Shivanshu Arora (126000303)
Kritika Kurani (825000784)

MOTIVATION

Information Explosion

Trending topics on Twitter

15 hours ago	16 hours ago	17 hours ago
Bayern	Bayern	Bayern
#كلمتين_وبس	#PınarÜrünleriBoykot	Leicester
#PınarÜrünleriBoykot	Leicester	#PınarÜrünleriBoykot
Leicester	#كلمتين_وبس	#80MilyonKardeşizBiz
#حاجات_تزيد_الانثي_انوثه	#استشهاد_12ضابط_سعودي_باليمن	#استشهاد_12ضابط_سعودي_باليمن
#LuanNoMusicaBoa	#80MilyonKardeşizBiz	Hummels
#ALlorarEnMTVHits	#FirstDates	Aytekin
Fresno	جمال الشريف	#WeLoveYouCamila
جمال الشريف	Hummels	#67حياتك
Aytekin	Fresno	Fresno

As of April 19th, 2017

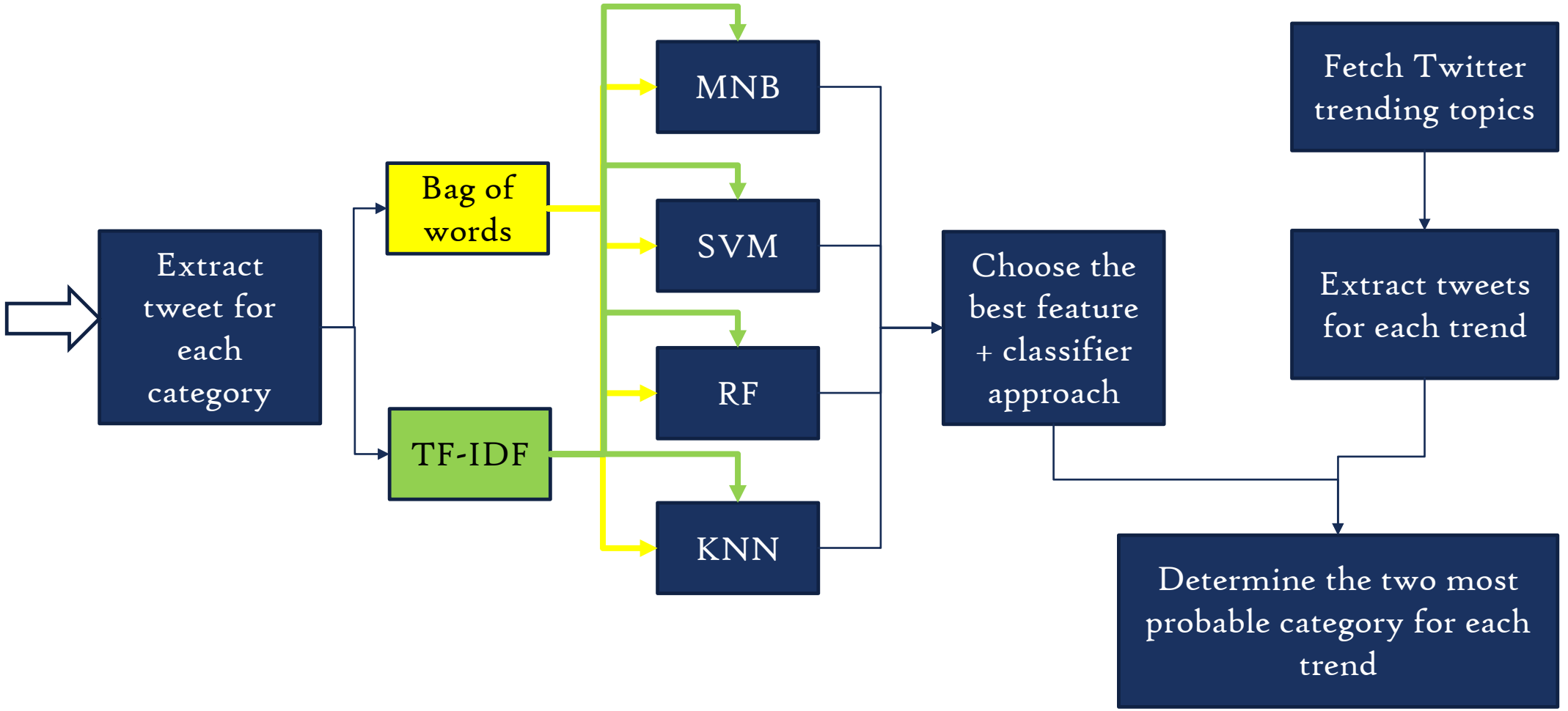
A trend on Twitter refers to a hashtag-driven topic that is immediately popular at a particular time.

MOTIVATION

Want to learn about a trend but what category does it belong to?

OUR APPROACH

- agriculture
- arts
- construction
- consumer goods
- corporate
- educational
- finance
- government
- high tech
- legal
- manufacturing
- media
- medical
- nonprofit
- recreational
- service
- transportation



DATASET CREATION

- agriculture
- arts
- construction
- consumer goods
- corporate
- educational
- finance
- government
- high tech
- legal
- manufacturing
- media
- medical
- nonprofit
- recreational
- service
- transportation

Hand curated keywords for each category

agriculture
- dairy
- farming

construction
- architecture and planning
- building materials
- civil engineering
- construction

educational
- e-learning
- education management
- higher education
- primary
- research

Fetch around 1000 tweets for each category, equally distributed among all keywords.

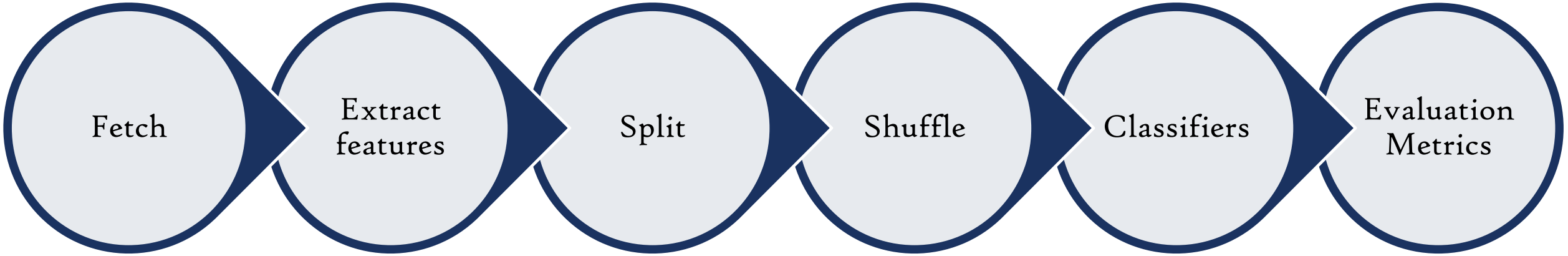
Normalize tweet data:

- Remove user names
- Remove URLs
- Remove punctuations
- Remove extra whitespaces
- Remove accents
- Remove stop words
- Convert to lowercase

Each tweet was labeled with it's general category name.

Tweets were saved to each category file

CLASSIFICATION OF TWEETS



- Fetch tweets from the category files

- Bag of Words
- TFIDF

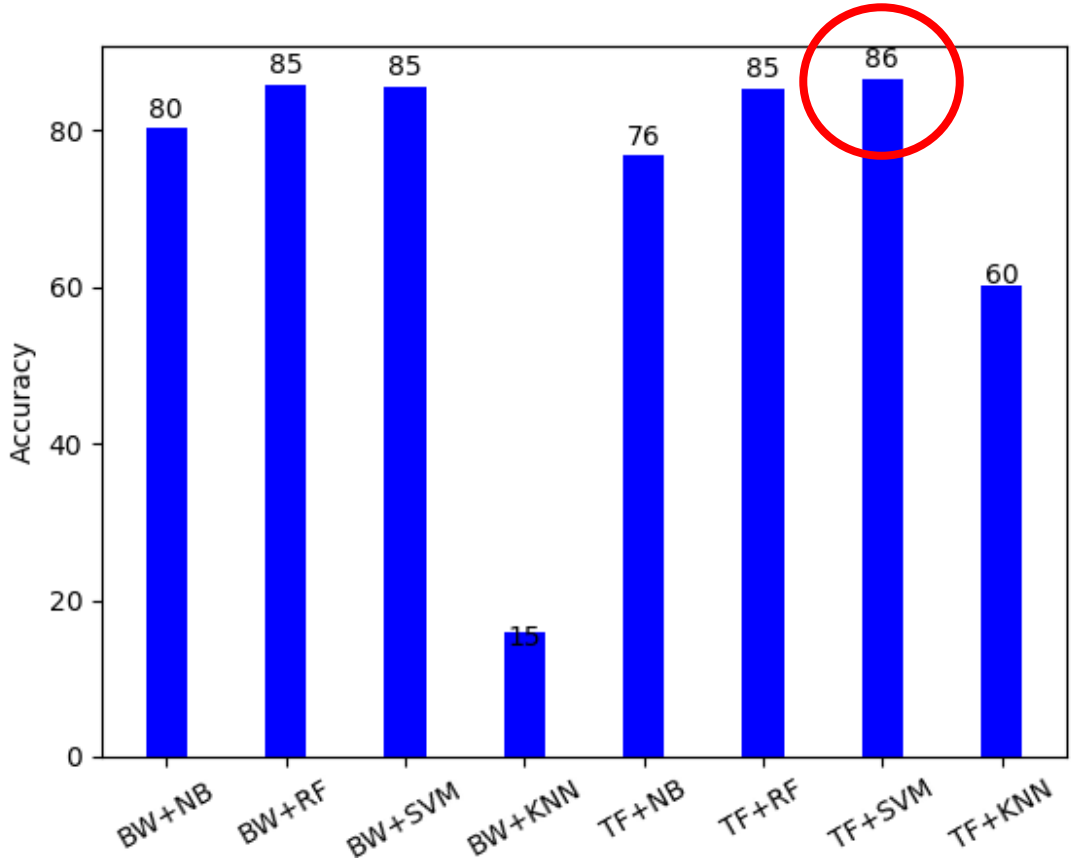
- Split the data from each category into training and testing sets
- 60:40

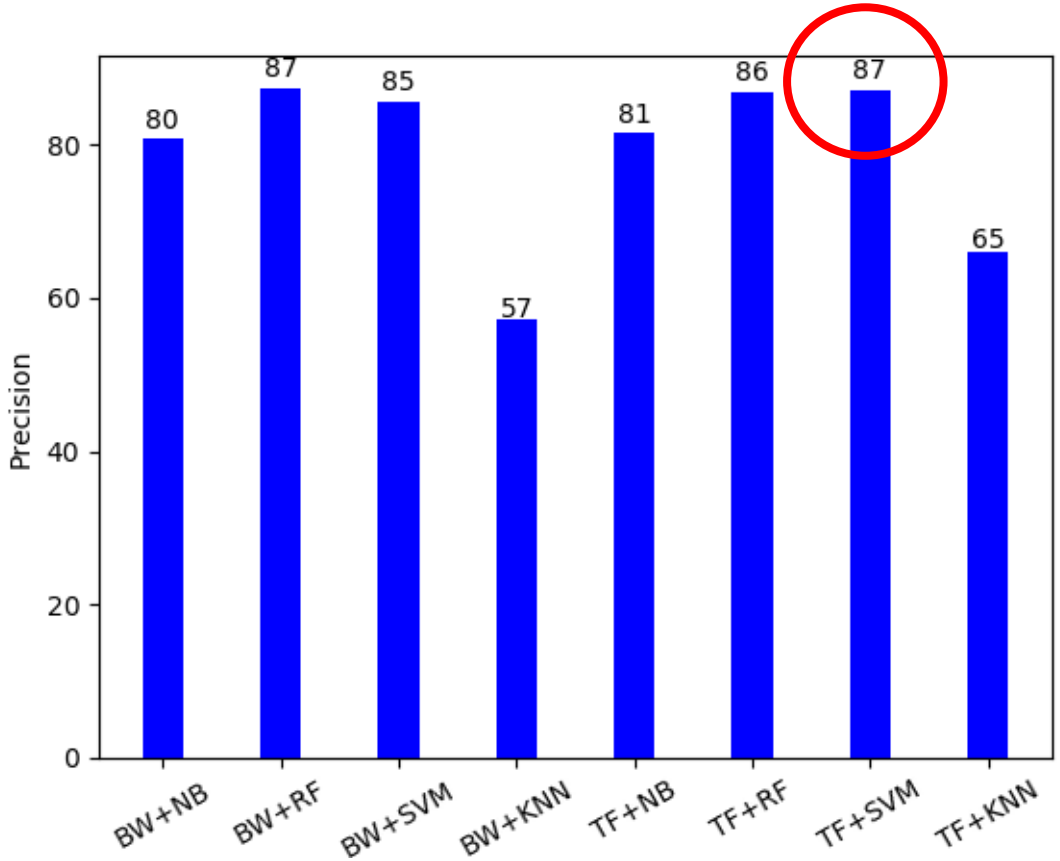
- Get rid of minibatches of highly correlated examples

- Multinomial Naïve Bayes
- Support Vector Machine
- Random Forest
- KNN

- Accuracy
- Precision
- F1

RESULTS OF CLASSIFICATION

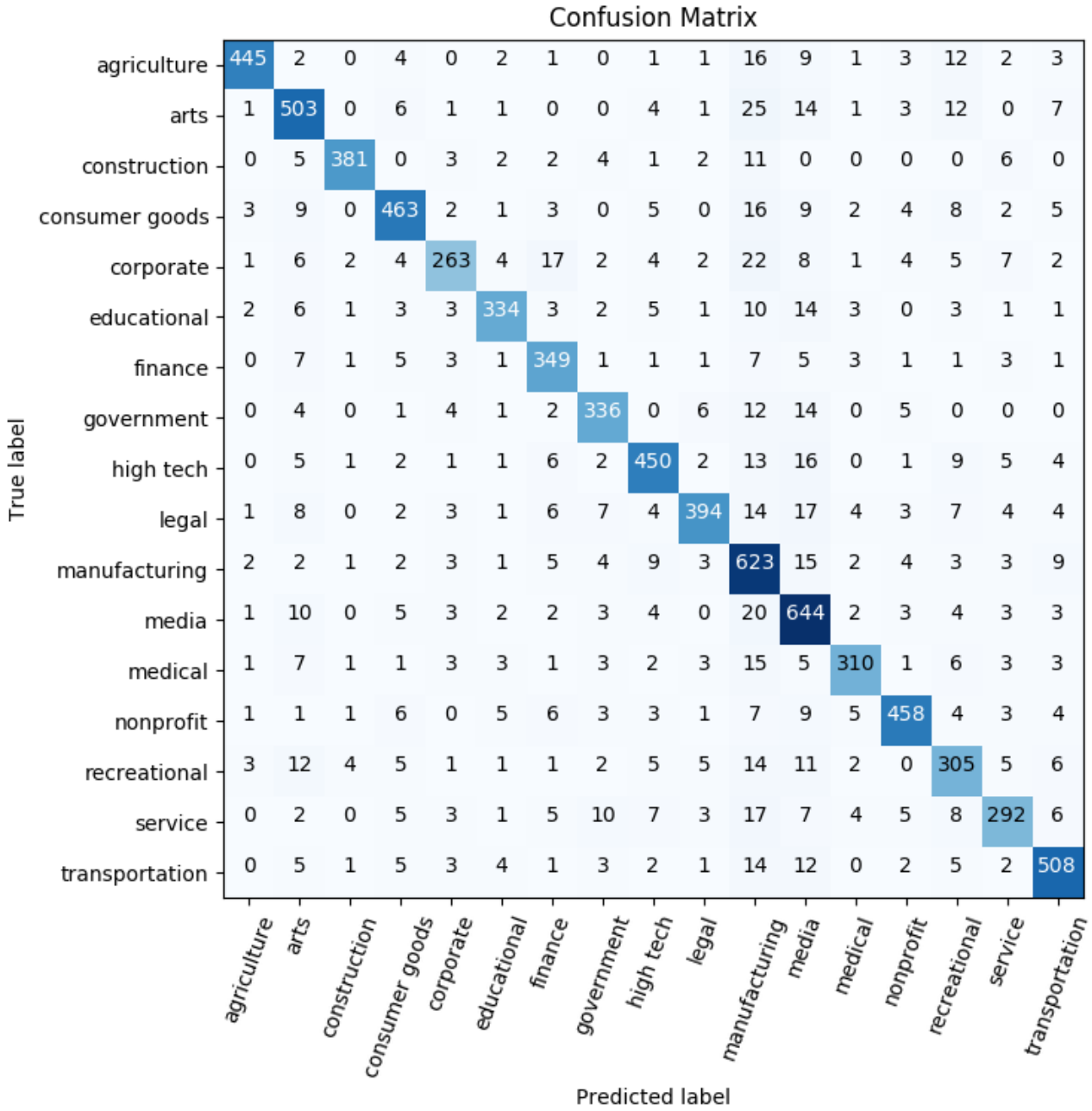




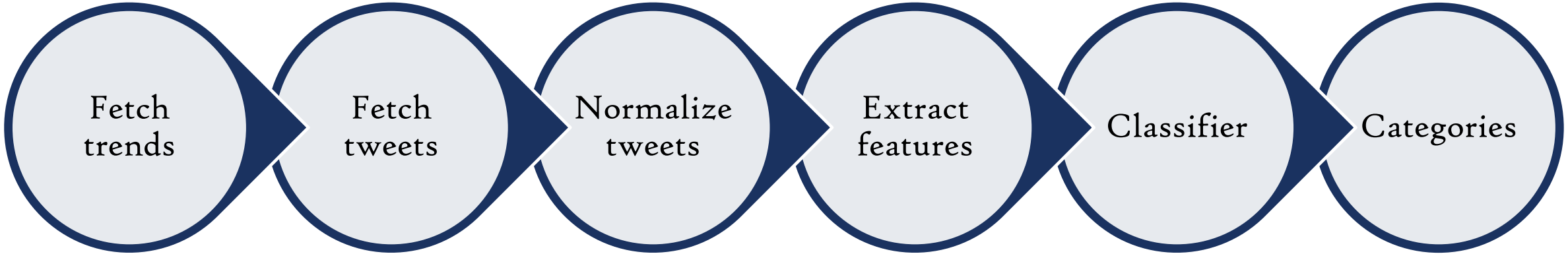
RESULTS OF CLASSIFICATION

Highest misclassified data to manufacturing and media categories

- Larger training set
- More correlated with other categories



TRENDING TOPIC CLASSIFICATION



- Trending topics using Twitter module
- For each trend, fetch around 100 tweets
- TF-IDF
- SVM
- Category label for each tweet
- Determine the two most probable categories

RESULTS OF TRENDING TOPIC CLASSIFICATION

WeLoveYouCamila
arts 0.36
service 0.17

CC1 🌹 @Iamwithccabello · 8h
Thank you for being the sweetest with your fans. We appreciate it so much.
#WeLoveYouCamila

Thalita. @TtroopCamilaC · 6h
"I know that no matter what happens, I am following my heart" - Camila Cabello
#WeLoveYouCamila

Xabi Alonso
recreational 0.45
consumer goods 0.21

Sara ~ 11 @iZubi_Sara · 3h
At least your last game was at the Bernabeu ! Thank you @XabiAlonso , you will be missed forever. ❤️ #XabiAlonso

Ricky Machel @rickymachel · 5h
What a guy. Deserved his standing ovation after the game. Pure class and a massive loss #ChampionsLeague nights #XabiAlonso

Fresno
media 0.45
government 0.2

Amir MC Spice Shakir @AmirQShakir · 2h
The killer in #Fresno is an idiot.Jesus NOR Allah approves of his actions. God bless the families of the victims.

Razor @hale_razor · 6h
Three killed... OH NO
..by gunman.. GUN CONTROL NOW!
..yelling Allahu Akbar. LETS DISCUSS TRUMPS TAXES

#Fresno

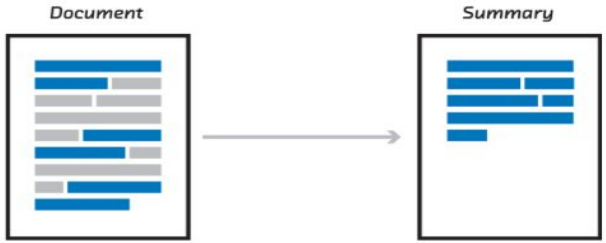

GRACIAS
THANK
YOU
ARIGATO
SHUKURIA
GOZAIMASHITA
EFCHARISTO
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EXHMET
MEHRBANI
GRAZIE
MAAKE
KOMAPSUMNIDA
JUSPAXAR
PALDIES
BOLZÏN
MERCI
DANKSCHEEN
BÏYAN
SHUKRIA
TINGKI

PRÉCIS

A Summarization Tool

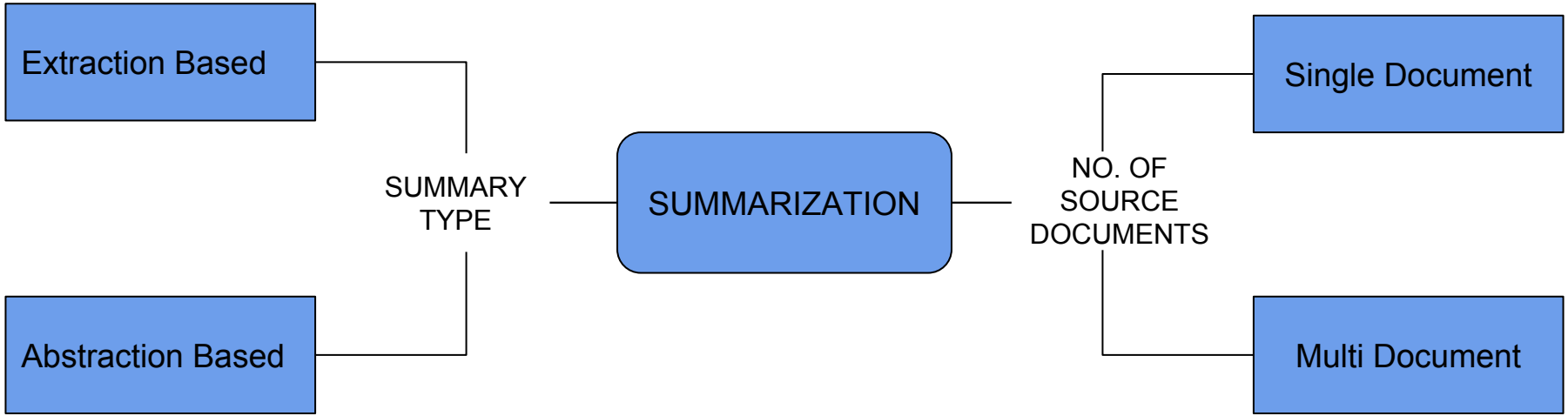
Abhishek Sharma, Shubham Jain

PROBLEM DEFINITION



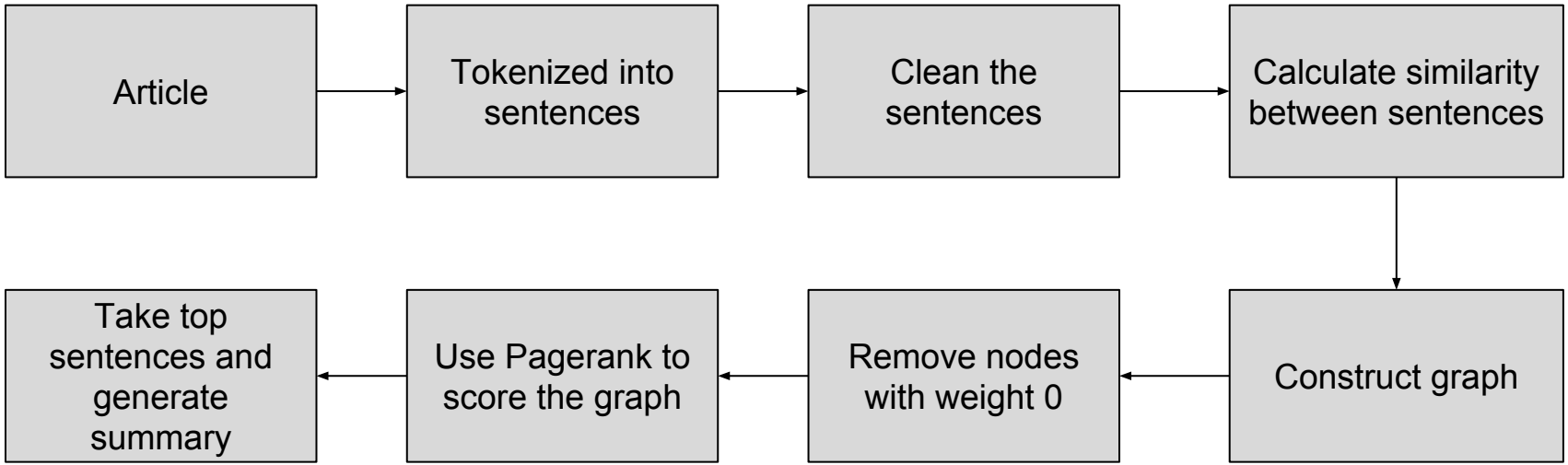
- With the increasing amount of digital information, it has become difficult for the reader to consume lot of things. Thus it is necessary to build a system that could produce human quality summaries.
- This would save a lot of time spent by humans on reading and thus, increase their efficiency.
- **Precis** is a tool that provides **summaries** of a given document.
- In this project, we have implemented the **Textrank** algorithm for text summarization.
- Also, we have created an interactive website and chrome extension for anyone to use.

CLASSIFICATION OF SUMMARIZATION TASKS



ALGORITHM USED - TEXTRANK

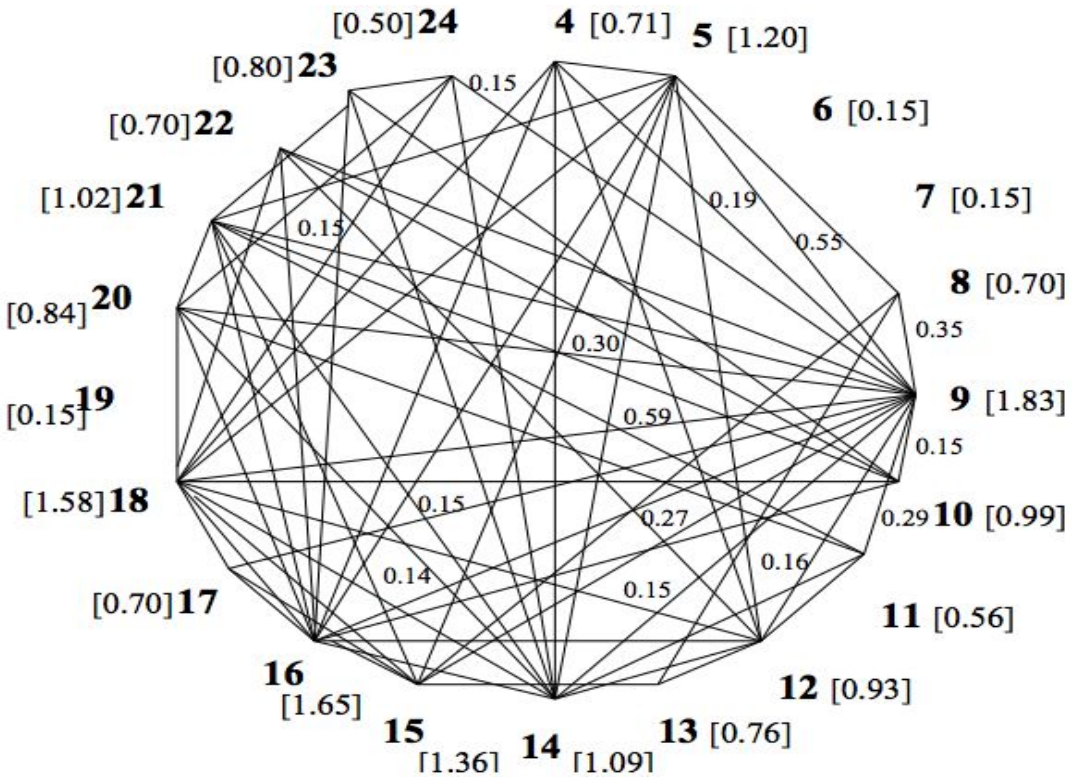
How Text Rank Algorithm Works?



ALGORITHM USED - TEXTRANK (EXAMPLE)

3: BC–HurricaineGilbert, 09–11 339
4: BC–Hurricaine Gilbert, 0348
5: Hurricaine Gilbert heads toward Dominican Coast
6: By Ruddy Gonzalez
7: Associated Press Writer
8: Santo Domingo, Dominican Republic (AP)
9: Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12: Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14. Tropical storm Gilbert formed in the eastern Carribean and strenghtened into a hurricaine Saturday night.
15: The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
19: There were no reports on casualties.
20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24: The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

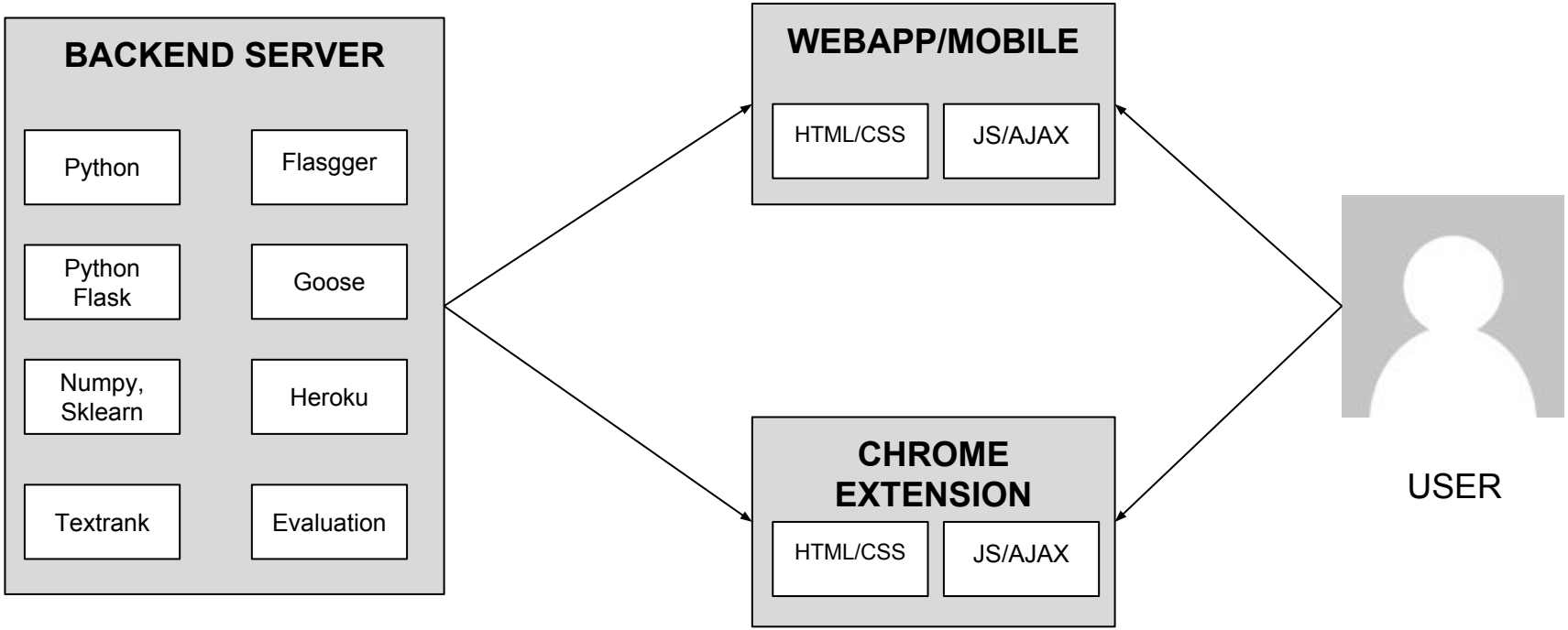
ALGORITHM USED - TEXTRANK (EXAMPLE)



ALGORITHM USED - TEXTRANK (EXAMPLE)

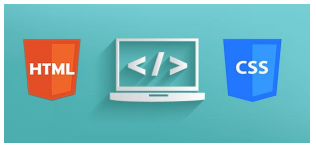
Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil De-
fense alerted its heavily populated south coast to prepare for high winds, heavy rains
and high seas. The National Hurricane Center in Miami reported its position at 2 a.m.
Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce,
Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service
in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad
area of cloudiness and heavy weather" rotating around the center of the storm. Strong
winds associated with Gilbert brought coastal flooding, strong southeast winds and up
to 12 feet to Puerto Rico's south coast.

Architecture of Precis



TECHNOLOGY STACK USED

- Language Used
  - Python (Backend Language)
  - HTML, Javascript, Ajax, CSS (Frontend Language)
  - Bash (Shell Scripting)
- Libraries Used
  - Numpy, Matplotlib
  - Sklearn, Goose, Flasgger, etc.
- Framework Used
  - Python Flask (API resource Layer)
- Deployment Environment
  - Herokuapp, Amazon AWS
- Development Environment
  - JetBrains PyCharm, Github

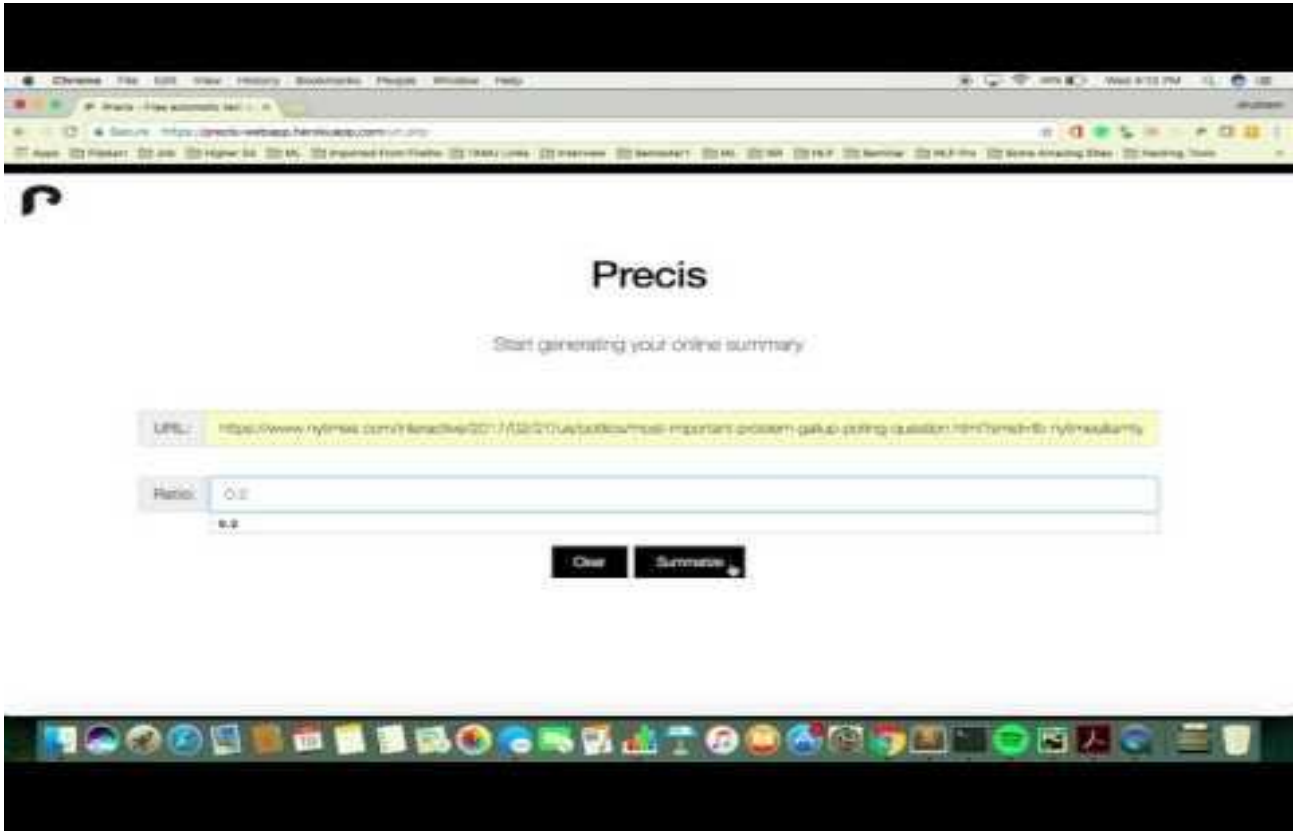



EVALUATION

- MultiLing 2015 Dataset: 30 documents, including text and human generated summaries
- Evaluation Toolkit used: ROUGE-1
- Results:

ALGORITHMS	**SCORE**
Text-Rank (Our algorithm)	0.354453733333
edmundson	0.300956866667
sum-basic	0.314202633333
lex-rank	0.327850166667

DEMO


Precis
Start generating your online summary
URL:
Ratio:
Clear
Summarize

IMPORTANT LINKS

Github Repository Link of the whole code:-
https://github.com/shubham7jain/precis

Demo
https://www.youtube.com/watch?v=K-l5UQ3vARE

Website Link
https://precis-webapp.herokuapp.com/

API Contract is available at
http://precis.herokuapp.com/apidocs/index.html

ARTICLE STRUCTURE IN NEWS STORIES

Himanshu Taneja
UIN: 426000238

NEWS STORIES

News is defined as:

“presentation of information about current events”

ARTICLE STRUCTURE

What is **article structure**?

In news stories, article structure is characterized by 2 key things:

- The order in which events are introduced and presented
- The writing style

ORDER?

Order is determined by **when** the **key event** of a story is introduced.

American doctor devises creative ways to help Haitians

She stared blankly into the far distance and moved in a robotic, shuffling gait as she was led by a neighbor into a free health clinic run by Dr. Bob Paeglow.

"Bon jour," Paeglow said.

His French greeting was met with silence. The woman was pale and thin.

Maryland women charged with hate crime after Trump sign burned

Two Maryland women have been charged with a hate crime after being accused of burning a "make America great again" sign.

Princess Anne police charged D'Asia R. Perry and Joy M. Shuford with second-degree arson and committing a hate crime, the Baltimore Sun reported.

WRITING STYLE?

Characterized by how the author is presenting events:

- Is he reporting facts?

  Or

- Is he narrating a story?

WRITING STYLE?

American doctor devises creative ways to help Haitians

She stared blankly into the far distance and moved in a robotic, shuffling gait as she was led by a neighbor into a free health clinic run by Dr. Bob Paeglow.

"Bon jour," Paeglow said.

His French greeting was met with silence. The woman was pale and thin.

Maryland women charged with hate crime after Trump sign burned

Two Maryland women have been charged with a hate crime after being accused of burning a "make America great again" sign.

Princess Anne police charged D'Asia R. Perry and Joy M. Shuford with second-degree arson and committing a hate crime, the Baltimore Sun reported.

WHY ARTICLE STRUCTURE?

Why should we care about the **order** and the **writing style**?

It can help us design algorithms that can target different type of news stories differently.

For example:

News Summarization

If we know the **order**, we know which region to target.

TYPES OF ARTICLE STRUCTURE

Structure in news stories can be broadly classified into 4 types:

- Inverted Pyramid
- Kabob
- Martini Glass
- Narrative
- Other, for the ones that don’t follow any of the above

INVERTED PYRAMID

Present the **most important/ relevant** events **first**. Followed by other information about these events.

- Most followed article structure
- Best suited for: **news briefs, breaking news**

MARTINI GLASS

Similar to Inverted Pyramid, except in the end it also presents a **narrative describing the key event** in detail.

- Best suited for: **crime**, **sports stories**, where author wants to:
  - first briefly summarize the key event, and
  - then provide a detailed narration of how the key event happened, step-by-step.

MARTINI GLASS

Shots Fired While He Stabbed Ex-Wife

….

Dennis Leach became angry with his 37-year-old ex-wife after he went to a neighborhood bar Friday night. He stormed into her duplex Saturday afternoon and threatened her with a butcher knife. A terrified Joyce Leach dashed next door to the adjoining home of Leach's parents. "He's got a knife, and he's gonna kill me!" Leach's mother, Reba Leach, said her daughter-in-law screamed. At the same time, 15-year-old April Leach, one of their six chil- dren, called from a convenience store blocks away. "Your father is going to kill me!" Joyce Leach yelled. April Leach hung up and dialed 911.

….|

KABOB

The story **begins with a anecdote** about a specific person/ thing. Then it **broadens into a general discussion** of the topic. It **ends by returning to that specific person/ thing** again.

- Second most followed article structure
- Best suited for stories where author may want to show how actual people are affected or involved. For example, to talk about *financial crisis & recession*, author would:
  - first describe a person facing hardships due to recession
  - then provide details about the *recession/ crisis*

NARRATIVE

The whole news story is **a narration of some events**.

- Best suited for articles where author wants to:
  - describe the event/ person in a very detailed manner
  - add a fictional touch to his story; making it more interesting to read

DATASET

English Gigaword

Distributed by **Linguistic Data Consortium**

For our task, we've used articles published by **NYTimes** during the year 2010

Stories were sampled from different categories:

- Business
- Crime
- Politics
- Disaster

GUIDELINES & ANNOTATION

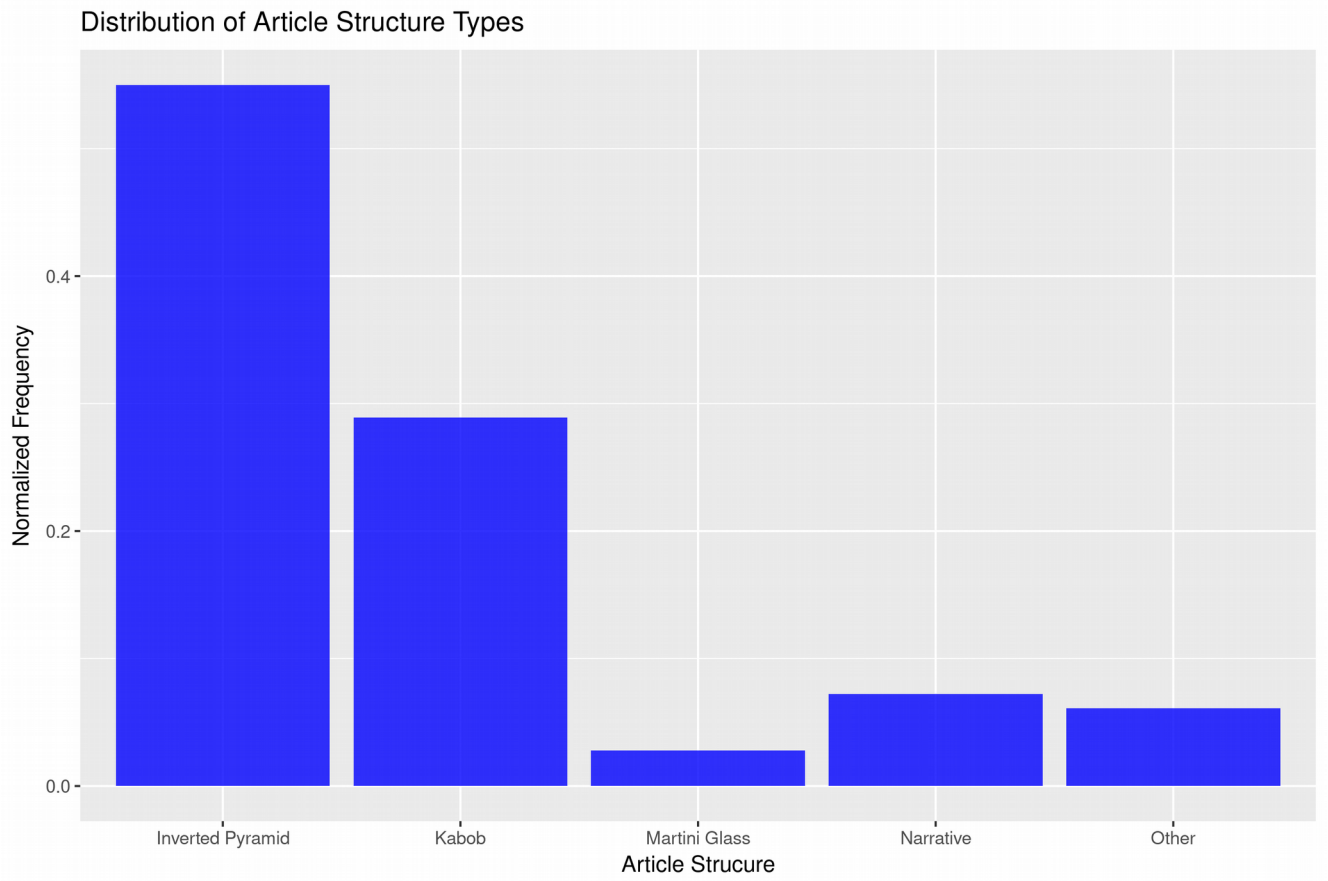
Guidelines for annotation were designed.

A total of 180 news stories (45 from each each category) were labeled.

The whole dataset was then split into 2 sets:

- Training (129 news stories)
- Testing (51 news stories)

DISTRIBUTION



	Politics	Crime	Business	Disaster
Inverted Pyramid	21	20	34	24
Kabob	19	10	7	16
Martini Glass	0	3	0	2
Narrative	1	7	3	2
Other	4	5	1	1

Structure Type	Number of News Stories
Inverted Pyramid	99
Kabob	52
Martini Glass	5
Narrative	13
Other	11
Total	**180**

FEATURES USED

A total of 20 different features used.

Two types of features were designed:

- **To capture the order**
  - Similarity between different parts of the news stories
  - N-grams of the first sentence of different paragraphs
  - NER tags for Person, Location, and Time

FEATURE USED

- To **capture the writing style**
  - Length of story; number of paragraphs, number of words, length of first and last paragraph
  - Presence of passive speech in first few paragraphs
  - Frequency of the word “said”

MACHINE LEARNING MODELS

3 different models were tried:

- Linear Support Vector Machines
- Support Vector Machines with Gaussian Kernel
- Random Decision Forest

Hyper parameters for each of the model were tuned over a grid using 10-fold Cross Validation

Final model selection was also performed using these cross validation results.

Model was selected which achieved the highest value for *Micro Averaged F1 Score.*

RESULTS

Random Forest was selected as the final model.

```
RANDOM FOREST RESULTS
10 fold cv estimate of accuracy: 65.12
10 fold cv estimate of micro avg f1: 0.78
Accuracy on test set: 72.54

Precision, Recall, and F-1 score for each class on test set:
                  precision    recall  f1-score
inverted_pyramid       0.86      0.83      0.84
           kabob       0.55      0.73      0.63
   martini_glass       0.00      0.00      0.00
       narrative       1.00      0.33      0.50
           other       0.50      0.33      0.40
     micro avg         0.75      0.75      0.75
```

RESULTS

Confusion Matrix:

	Reference				
Prediction	invt_pyramid	kabob	mrtni_glss	narrative	other
inverted_pyramid	24	3	1	0	0
kabob	5	11	0	2	2
martini_glass	0	0	0	0	0
narrative	0	0	0	1	0
other	0	1	0	0	1

Seems to work well on the two most frequent classes: Inverted Pyramid & Kabob

Why so poor performance on the other classes?
Maybe we need more data.

Structure Type	**Number of News Stories**
Inverted Pyramid	99
Kabob	52
Martini Glass	5
Narrative	13
Other	11
Total	**180**

QUESTIONS?

Soccer Key Event Extraction

Rahul Ashok Bhagat

Motivation

- Soccer – One of the most watched sports in the world.
- Nearly played by 250 million players in over 200 countries.
- With the hectic schedule of working class people, no time to watch match. What to do?
- Highlights – Video , Audio and Text

- Soccer Analytics
- Many events involved in soccer – Difficult compared to other sports.

Soccer Analytics

Previous Works

- Extract highlights automatically using audio-track features.
- Visual Analytics techniques into the analysis process.
- Game-related performance of the players and teams.
- Predicting Soccer matches (Betting).
- Mobile application usage for real-time opinion sharing.
- Social Media Data for event extraction.

Data Collection

- Various available sources:
  - Goal
  - Twitter
  - BBC Sports

- Goal.com – Primary source
  - 500+ strong editorial team
  - available in 18 languages across 38 location-based editions

Commentary Data

- Contains most information in text format.
- Good source for text highlights.
- Beneficial for Blind and Deaf People.

32'		Assist Łukasz Teodorczyk Teodorczyk collects the rebound and bundles the ball to Hanni for the midfielder to strike.
32'	⚽	**Goal Sofiane Hanni** **GOALLLL!!!! HUGE MOMENT FOR ANDERLECHT! Hanni finds the back of the net to score a crucial away goal for the Belgian outfit, hammering the ball past Romero from close range. Tielemans caused havoc in the box as his deflected strike hit the bar, while Teodorczyk did just enough to steer the ball into Hanni's path to strike with power into the net.**
30'		Shaw makes a surge down the left flank and sends a dangerous ball across the face of goal, but Spajic gets just enough on his clearance to take it away from Lingard in the box.
29'		SAVE! Tielemans curls the resulting free-kick over the wall, but Romero is on hand to make a comfortable save.
28'		Hanni has his ankles clipped by Pogba and the visitors have a free-kick in a dangerous position.

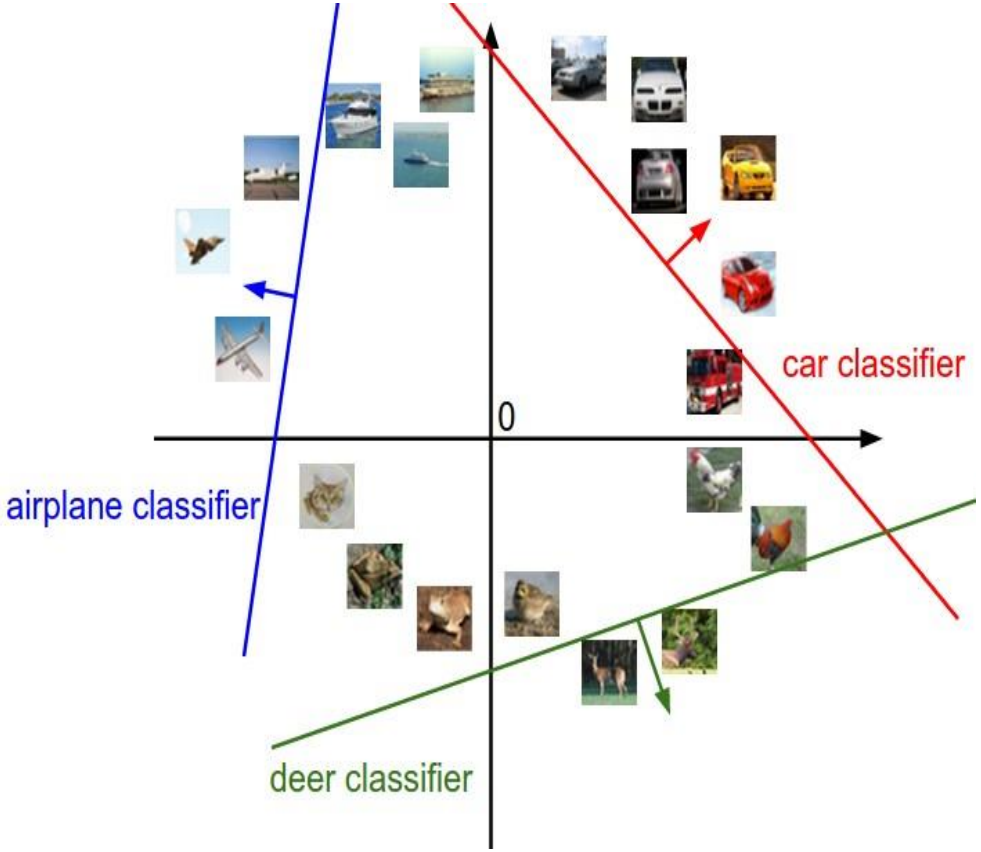
Event Types

- Goal
- Substitution
- Assist
- Yellow Card
- Red Card
- Penalty goal
- Penalty Save

  and many more

Event Classification

- 2 Classifiers:
- Binary Classifier – ‘Action – Not Just Action’
- Event Classifier – Goal, Assist, Substitution….



Different Models

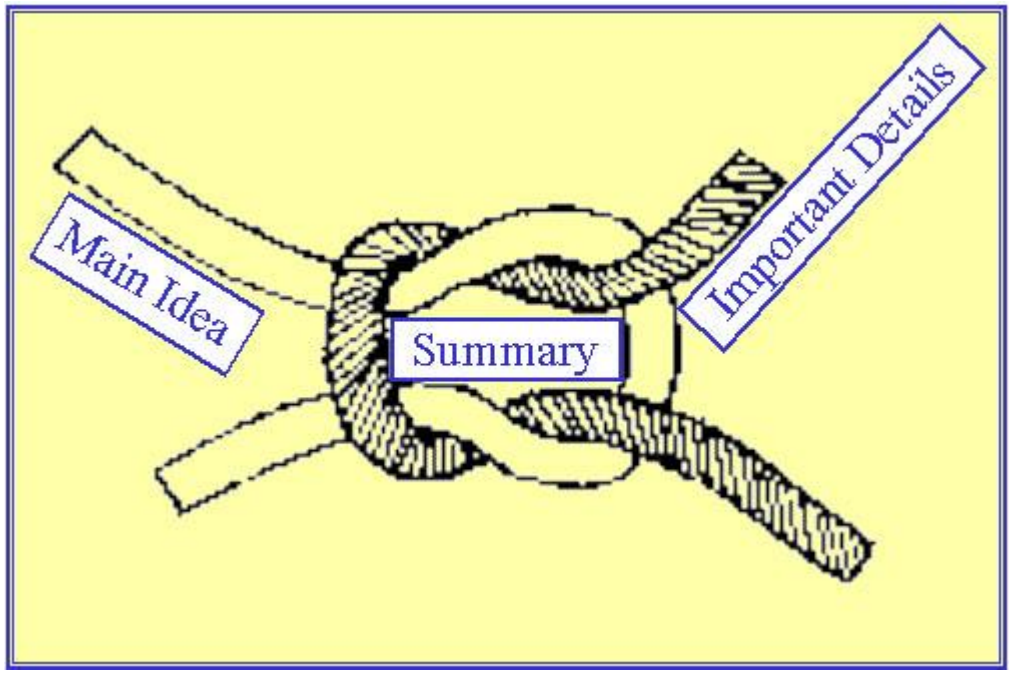
- Naïve Bayes
  - Gaussian
  - Bernoulli
  - Multinomial
- MaxEnt
  - Logistic Regression

	Precision	Recall	F1-Score
Action	0.97	0.92	0.94
Not Just Action	0.98	0.99	0.98
Avg/Total	0.97	0.97	0.97

	Substitution	Yellow Card	Assist	Goal	Penalty-goal	Own-goal	Yellow-red	Red-card	Penalty-save	Missed-penalty
Substitution	1045	0	0	0	0	0	0	0	0	0
Yellow card	0	721	1	3	0	0	0	0	0	0
Goal	0	0	509	4	0	0	0	0	0	0
Assist	0	5	7	381	0	0	0	0	0	0
Penalty - goal	0	0	9	0	18	0	0	0	0	0
Own-goal	0	0	10	0	0	6	0	0	0	0
Yellow-red	0	6	1	0	0	0	8	0	0	0
Red-card	0	1	2	0	0	0	2	3	0	0
Penalty-save	0	1	0	0	0	0	0	0	5	0
Missed-penalty	0	0	1	0	3	0	0	0	0	6

Text Summarization

- TextRank
- PageRank
- Latent Semantic Analysis



Result

Commentary 2: *Substitution sub-out Idrissa Gana Gueye sub-in Enner Remberto Valencia Lastra . Valencia replaces Gueye, as Koeman looks to take the game to Burnley a little. Barkley moves infield as a result.*

TextRank: *Valencia replaces Gueye as Koeman looks to take the game to Burnley a little.*

LexRank: *Substitution sub-out Idrissa Gana Gueye sub-in Enner Remberto Valencia Lastra .*

Future Work

- ’Saves’, ’Blocks’ and ’Chances’
- Rarely Occuring Events – Red Card , Penalty Missed etc
- Evaluating the player performance

QUESTIONS

NLP Final Project

Multi-Pass Sieve for Coreference Resolution

By

Jigna Reshamwala and Abhipsa Misra

Coreference Resolution

- Identify all noun phrases mentions) that refer to the same real world entity

Types of Coreferences

Anaphora

- The **music** was so loud that **it** couldn't be enjoyed.

Cataphora

- If **they** are angry about the music, the **neighbors** will call the cops.

Split antecedents

- **Carol** told **Bob** to attend the party. **They** arrived together.

Coreferring noun phrases.

- **Some of our colleagues** are going to be supportive. **These kinds of people** will earn our gratitude.

Our Approach

- Input- Tagged mentions, each with an id
- Multi-Pass Sieve- The mentions are matched with another id
- Evaluation- Precision- # of correct matches/# of pairs resolved
  Recall- # of correct matched/# of total pairs to be resolved

Features

- Incremental approach
- Rule based “unsupervised”
- Deterministic
- Multiple passes over text
- Precision of each pass is lesser than the preceding passes
- Recall keeps increasing with each pass
- Decisions once made cannot be modified by laterpasses
- Rule based “unsupervised”

Multi-Pass Sieves

Exact Nominal Phrase Matching

Precise Constructs- Acronyms and Appositives

Head Matching

Lexical Matching of Nominal Phrases

Pronouns Matching

Evaluation Metric

- Precision- # of correct matches/# of pairs resolved
- Recall- # of correct matched/# of total pairs to be resolved

Trends Observed

- Precision decreases with each pass
- Recall increases with each pass

Results

Passes	Precision	Recall
{1}	0.8773	0.3886
{1,2}	0.7640	0.4443
{1,2,3}	0.6908	0.5435
{1,2,3,4}	0.6894	0.5489
{1,2,3,4,5}	0.6656	0.5571

Table: Cumulative performance on development as passes are added to the sieve.

Conclusion

- Competitive end-to end coreference resolution system can be built using only deterministic models (or sieves).
- These models incorporate lexical, syntactic, and semantic information
- Our results demonstrate that, despite their simplicity, deterministic models for coreference resolution obtain competitive results
- We obtained comparable results to [1] and [2].

Future Scope

- Getting information on parse tree structure can improve pronoun sieve
- Addition of sieves on speaker identification and alias detection
- Addition of sieves that can detect well cataphora and split antecedents

References

1. Raghunathan, Karthik, et al. "A multi-pass sieve for coreference resolution." Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.
2. Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011.
3. Lee, Heeyoung, et al. "Deterministic coreference resolution based on entity-centric, precision-ranked rules." Computational Linguistics 39.4 (2013): 885-916.
4. Rao, Delip, Paul McNamee, and Mark Dredze. "Entity linking: Finding extracted entities in a knowledge base." Multi-source, multilingual information extraction and summarization. Springer Berlin Heidelberg, 2013. 93-115.
5. Recasens, Marta, Marie-Catherine de Marneffe, and Christopher Potts. "The Life and Death of Discourse Entities: Identifying Singleton Mentions." HLT-NAACL. 2013.

Neural Networks Models for Language Modeling and Text Similarity

Lingyiqing Zhou

Contents

- Tasks and Problems
- High-Level Description of Approach
- Specific System Implementation
- Evaluation Data, Metric and Experimental Results
- Conclusion
- Reference

Tasks and Problems

- Train a character-level language model by using the Recurrent Neural Networks(RNNs) models : the Long-Short Term Memory(LSTMs) models.

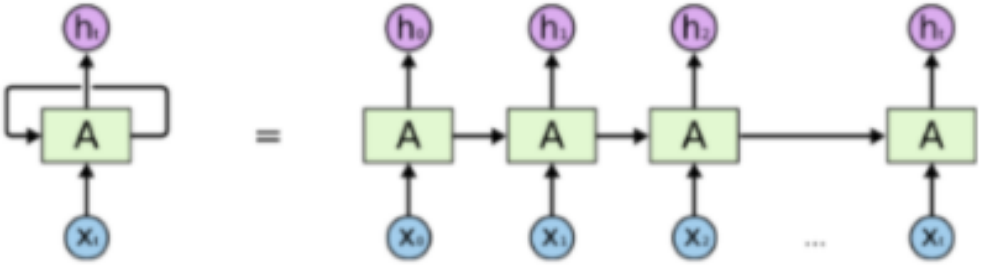
- Compare the results with the rest part of the input texts.

Description of the Approach

- Brief Introduction of the Neural Networks Models
  - RNNs(vanilla)
  - Long-Short Term Memory(LSTMs)

Description of the Approach

- RNNs



An unrolled recurrent neural network.

The main feature of the process in training RNNs is that there is backpropagation in time (BackPropagation Through Time, BPTT).

Description of the Approach

- Brief Mathematical derivation of the RNNs
  - Forward Propagation

$$a_k^t = \sum_{h=1}^{H} b_h^t \omega_{hk}$$

$$a_h^t = \sum_{i=1}^{I} x_i^t \omega_{ih} + \sum_{h'=1}^{H} b_{h'}^{t-1} \omega_{hh'}$$

$$b_h^t = \theta_h(a_h^t)$$

  - Back Propagation

$$\delta_h^t = \theta'(a_h^t)\left(\sum_{k=1}^{K} \delta_h^t \omega_{hk} + \sum_{h'=1}^{H} \delta_{h'}^{t+1} \omega_{hh'}\right)$$

  - Gradient Solution

$$\frac{\partial L}{\partial w_{ij}} = \sum_{t=1}^{T} b_i^t \delta_j^t$$

The subscript i,h,k means the input layer, the hidden layer and the output layer. a means the inactive value and b means active value.

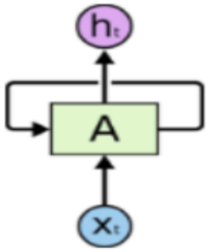
Description of the Approach

- Shortcoming: diffusion of the gradient.
  - With the increase of the network, the scope of the gradient would rapidly decrease.
  - The weights of derivate of the first levels are pretty small in the entire loss function.

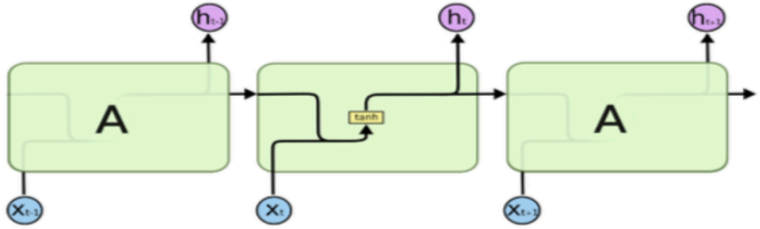
- Result: weights of some levels are renewed in a very low speed.

Description of the Approach

- LSTM



RNN

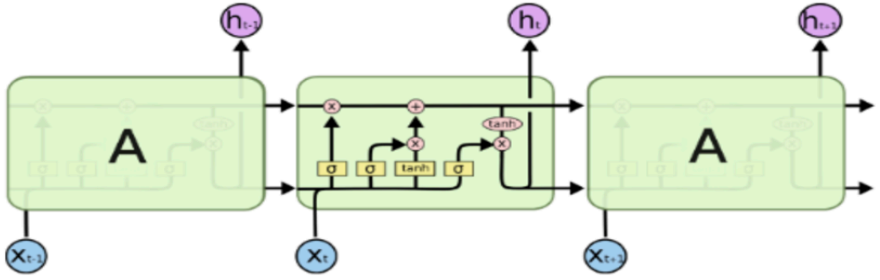


The repeating module in a standard RNN contains a single layer.

Vanilla RNN Unit

The biggest difference is the cell state.

The cell state runs straight down the entire chain.



The repeating module in an LSTM contains four interacting layers.

LSTM Unit

The information could stay in an unchanged state.

Description of the Approach

- Other Important Tools
  - Tensorflow
    - Open-source software library for numerical computation using data flow graphs;
    - Many frameworks such as Sequence-to-Sequence model and Language Modeling mode.
  - Keras
    - High-level neural networks API written in Python;
    - Capable of running on top of TensorFlow and Theano;
    - Model: core data structure of Keras, a way to organize layers.

Description of the Approach

- Other Important Tools
  - Gensim
    - Python library for topic modeling, document indexing and similarity retrieval with large corpora.
    - Use NumPy, SciPy and optionally Cython for performance.
    - Include implementations of TF-IDF, random projections, word2vec and document2vec algorithm.

Specific System Implementation

- Language Modeling
  - Get the vectors
    - Split the corpus into sentences
      - » Set a max length of sentences
      - » Store the sentences in a list
    - Vectorize the words in the sentences
      - » Two matrixes to store the information in order to reduce the spaces

Specific System Implementation

- Language Modeling
  - Use Models in Keras
    - Build Models
      - » Use the Sequential model
        - Add different layers to the model
      - » Add the LSTM layers
        - Set 2 Layers
        - 64 Hidden Neurons
    - Train Models
      - » 10 Epochs

Specific System Implementation

- Similarity Comparison
  - PreProcessing
    - Remove the common words
    - Remove the words that appear only once
    - Convert all the characters into lowercase
    - Remove all the punctuation
  - TF-IDF
    - TF: Term Frequency
    - IDF: Inverse Document Frequency
    - Reflect the importance of a word in the document

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = log \frac{|D|}{|\{j: t_i \in d_j\}| + 1}$$

Specific System Implementation

- Similarity Comparison
  - Latent Semantic Indexing(LSI) and Document2vec
    - Recognize the second-order co-occurrence words whose units are document;
    - Classify the words above into the same subspace;
    - Singular Vector Decomposition (SVD) is a way of matrix decomposition;
    - The module of LSI in Gensim implements fast truncated SVD.
  - Cosine Similarity
    - Measure of similarity between two non-zero vectors of an inner product.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Evaluation Data and Results

- Language Modeling
  - The First epoch

```
------------- EPOCH1 ----------------
Time taken:
279.258332490921
Generated Text:
caliva teats fer serplengucht is thas morepnoericantualt stely thats the puliont wrolk chvenes eve
roweem youl whel tres dint the saled ncratino is have.
shes a thus withts deduam unity tokatamy. us uhaliguttily fyoul thats sty hant on oule nout in anss
abehtologing.
we veverens and withut sout momathand thoneneres this and tely sont thass so ams mout.
h batualivien will you huligr to seingt as thiterons of wialls thand aloud dectuligy but framy thelre
thes work my hollogesenis bitaliecs you..
y
```

  - The Last Epoch

```
Time taken:
276.06477522850037
Generated Text:
 i giant tunisia. we were telling me as a great listen. what about sponsion form  markets newarm i
condition.
laughter
thank you.
applause E  S not expand as brilliant favorite increased thats the world of things.
thank you.
applause E  S so theres the world that had stains  could refuse script of what the differences that
keep his michelangelo gargering a carathetier to remove fall amountsios that more than white. if we
were able to leave it and received more competence of . but has been experiencing mad i could go down
there.
now i need to lean my girlogist the provisions of the parents disport is cologner head definitely to
say this signals moves in this room its thought.
in the freeder which is a kind of reality. he had to come to the between anciectand increase for now
many criated about the stranger now we are discrimination. of everything impossible and not publishagos
me. all one whole free. and i wanted to share and i was sitries its interesting term dig. the internet
a circui
```

Evaluation Data and Results

- Language Modeling
  - The decreasing of spelling mistakes and the Grammar Mistakes;
  - Difficult to use the cosine similarity
    - The size of the output is much smaller than the input
    - Hard to get the rest of the input data
    - Spelling mistakes would influence the accuracy
  - Extract some “common words” and compare the frequency
    - Have a great relation with the choice of the word and the size
    - “to” 3.62% VS 7.6% 3.62% VS 4.75%
    - “from” 2.72% VS 5.41% 2.72% VS 3.1%

Evaluation Data and Results

- Document Similarity
  - The input :

```
#Read File
documents = ["Human machine interface for lab abc computer applications",
             "A survey of user opinion of computer system response time",
             "The EPS user interface management system",
             "System and human system engineering testing of EPS",
             "Relation of user perceived response time to error measurement",
             "The generation of random binary unordered trees",
             "The intersection graph of paths in trees",
             "Graph minors IV Widths of trees and well quasi ordering",
             "Graph minors A survey"]
```

```
new_doc = "Graph theory is important for Human"
```

  - The output

```
[(8, 0.95881504), (7, 0.87431633), (6, 0.860421), (5, 0.84122026), (1, 0.55335683), (4, 0
.49610272), (2, 0.45502615), (0, 0.45487368), (3, 0.45070273)]
```

The 8th text of the input document has the highest similarity.
The 3th text of the input document has the lowest similarity.

Evaluation Data and Results

- Document Similarity
  - Some tricks
    - In the preprocessing, there is no need to remove the "common words" such as the preposition in some cases
  - Reason
    - If the word is really "common", they would appear in the corpus for many times.
    - Their IDF would be 0.

Conclusion

- Function
  - Generate some texts whose style of writing is similar to the input document.
  - Present the similarity of some documents in a numerical way.
- Result
  - Less Mistakes
  - Closer Frequency
- Still need much improvement
  - Algorithm: LightRNN
  - Better Equipment

Reference

1.http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/
2.http://colah.github.io/posts/2015-08-Understanding-LSTMs/
3. http://www.voidcn.com/blog/dream_catcher_10/article/p-2421992.html
4.http://www.52nlp.cn/%E5%A6%82%E4%BD%95%E8%AE%A1%E7%AE%97%E4%B8%A4%E4%B8%AA%E6%96%87%E6%A1%A3%E7%9A%84%E7%9B%B8%E4%BC%BC%E5%BA%A6%E4%BA%8C
5.https://www.tensorflow.org/
6. http://cloga.info/python/2014/01/27/corpora_vector_space
7.https://zh.wikipedia.org/wiki/Tf-idf
8.https://lizrush.gitbooks.io/algorithms-for-webdevs-ebook/content/chapters/tf-idf.html
9.http://d0evi1.com/gensim/tut1/
10.http://radimrehurek.com/gensim/tut2.html