Natural Language Processing (NLP)

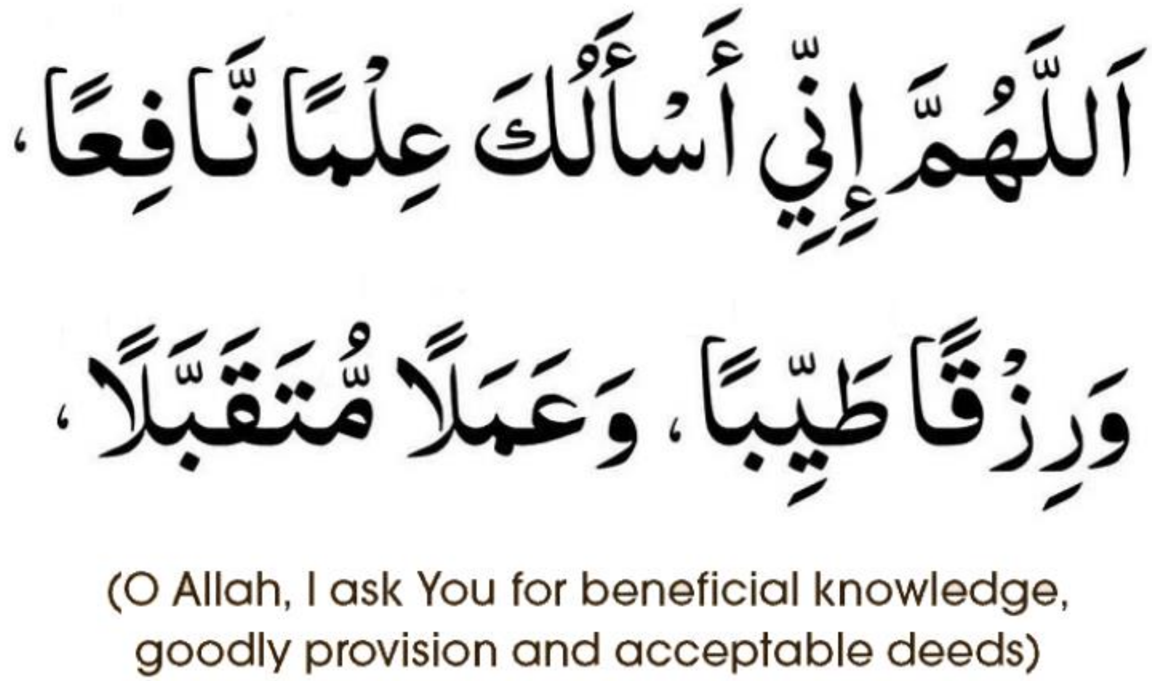
Unit 05

Text Vectorization

By:

Syeda Saleha Raza

AL NAFI,
A company with a focus on education,
wellbeing and renewable energy.



اَللَّهُمَّ إِنِّي أَسْأَلُكَ عِلْمًا نَّافِعًا،

وَرِزْقًا طَيِّبًا، وَعَمَلًا مُّتَقَبَّلًا،

(O Allah, I ask You for beneficial knowledge,
goodly provision and acceptable deeds)

اے اللہ ، میں آپ سے سوال کرتی ہوں نفع بخش علم کا، طیّب رزق کا، اور اس عمل کا
جو مقبول ہو.

(Sunan Ibn Majah: 925)

References

- CS-4650/7650: Natural Language Processing (gatech.edu)
- http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
- https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

Outline

- Text Classification
- Vectorizing Text
  - Bag of Words (BOW) ✓
  - TF-IDF ✓
  - N-gram ✓
  - Word Embeddings ✗

Text Classification

Movie Rating

positive

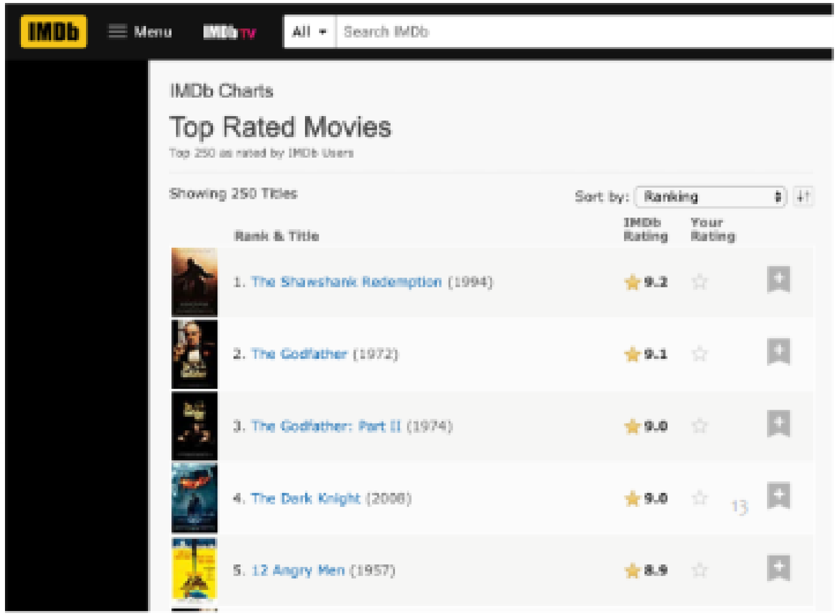
"… is a film which still causes real, not figurative, chills to run along my spine, and it is certainly the bravest and most ambitious fruit of Coppola's genius"

Roger Ebert, Apocalypse Now

- "I hated this movie. Hated hated hated hated hated this movie. Hated it. Hated every simpering stupid vacant audience-insulting moment of it. Hated the sensibility that thought anyone would like it."

Roger Ebert, North

negative



IMDb Charts

Top Rated Movies

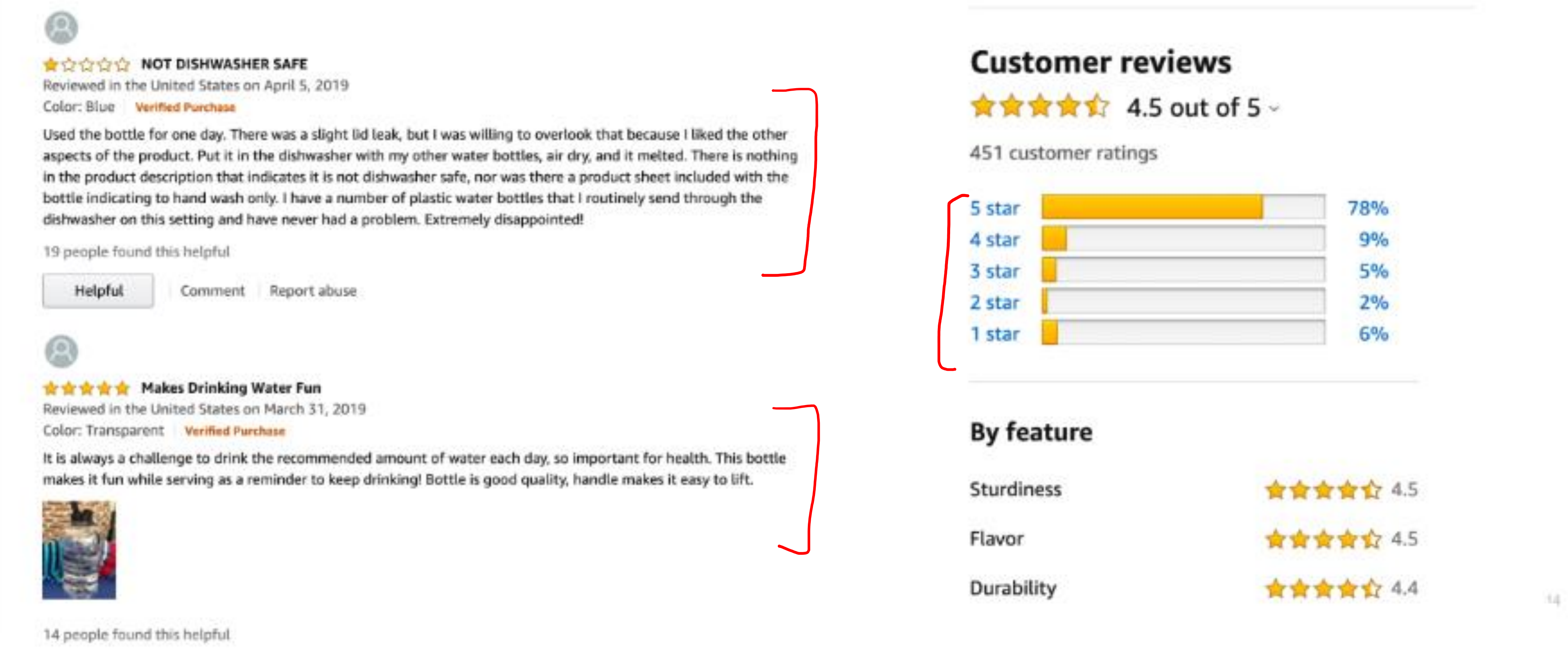
Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: Ranking

Rank & Title	IMDb Rating	Your Rating
1. The Shawshank Redemption (1994)	★ 9.2	☆
2. The Godfather (1972)	★ 9.1	☆
3. The Godfather: Part II (1974)	★ 9.0	☆
4. The Dark Knight (2008)	★ 9.0	☆
5. 12 Angry Men (1957)	★ 8.9	☆

Customer Reviews



★☆☆☆☆ **NOT DISHWASHER SAFE**

Reviewed in the United States on April 5, 2019

Color: Blue | Verified Purchase

Used the bottle for one day. There was a slight lid leak, but I was willing to overlook that because I liked the other aspects of the product. Put it in the dishwasher with my other water bottles, air dry, and it melted. There is nothing in the product description that indicates it is not dishwasher safe, nor was there a product sheet included with the bottle indicating to hand wash only. I have a number of plastic water bottles that I routinely send through the dishwasher on this setting and have never had a problem. Extremely disappointed!

19 people found this helpful

★★★★★ **Makes Drinking Water Fun**

Reviewed in the United States on March 31, 2019

Color: Transparent | Verified Purchase

It is always a challenge to drink the recommended amount of water each day, so important for health. This bottle makes it fun while serving as a reminder to keep drinking! Bottle is good quality, handle makes it easy to lift.

14 people found this helpful

Customer reviews

★★★★☆ 4.5 out of 5

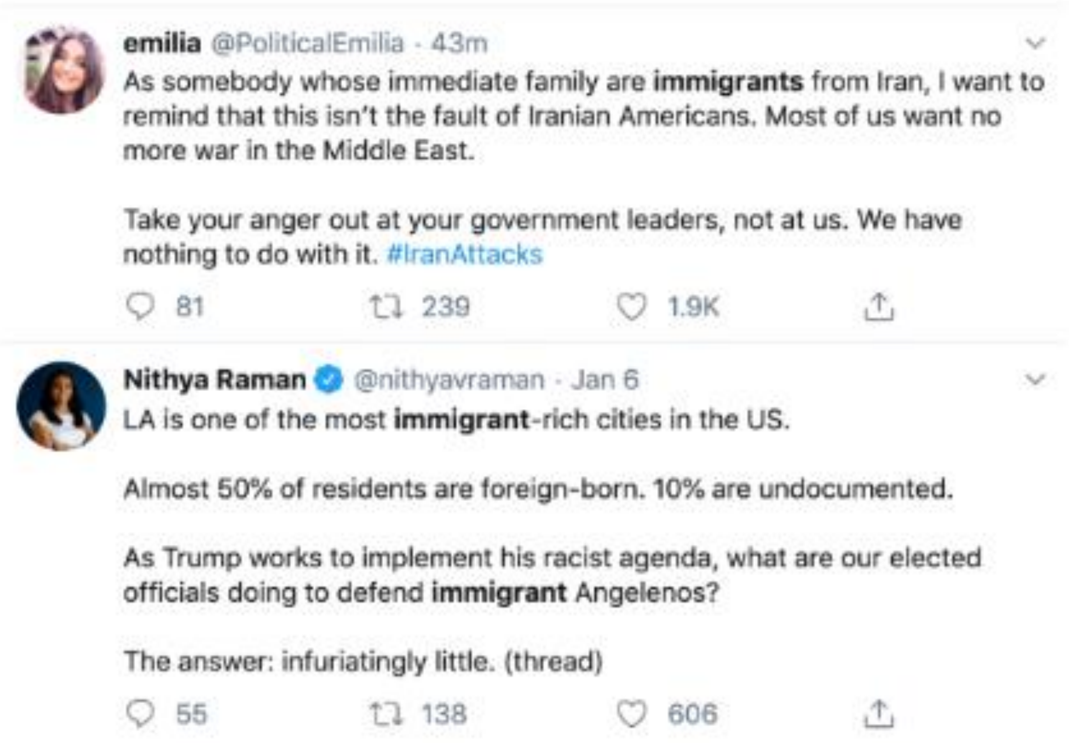
451 customer ratings

5 star	78%
4 star	9%
3 star	5%
2 star	2%
1 star	6%

By feature

Sturdiness	★★★★☆ 4.5
Flavor	★★★★☆ 4.5
Durability	★★★★☆ 4.4

Opinion Mining



emilia @PoliticalEmilia · 43m

As somebody whose immediate family are **immigrants** from Iran, I want to remind that this isn't the fault of Iranian Americans. Most of us want no more war in the Middle East.

Take your anger out at your government leaders, not at us. We have nothing to do with it. #IranAttacks

81 239 1.9K

Nithya Raman @nithyavraman · Jan 6

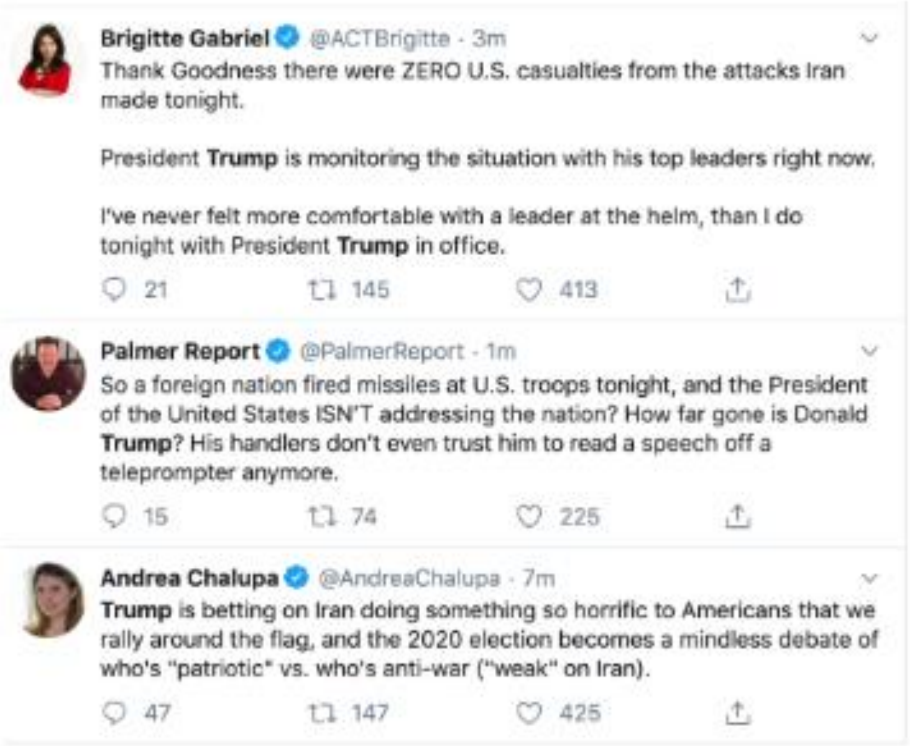
LA is one of the most **immigrant**-rich cities in the US.

Almost 50% of residents are foreign-born. 10% are undocumented.

As Trump works to implement his racist agenda, what are our elected officials doing to defend **immigrant** Angelenos?

The answer: infuriatingly little. (thread)

55 138 606



Brigitte Gabriel @ACTBrigitte · 3m

Thank Goodness there were ZERO U.S. casualties from the attacks Iran made tonight.

President **Trump** is monitoring the situation with his top leaders right now.

I've never felt more comfortable with a leader at the helm, than I do tonight with President **Trump** in office.

21 145 413

Palmer Report @PalmerReport · 1m

So a foreign nation fired missiles at U.S. troops tonight, and the President of the United States ISN'T addressing the nation? How far gone is Donald **Trump**? His handlers don't even trust him to read a speech off a teleprompter anymore.

15 74 225

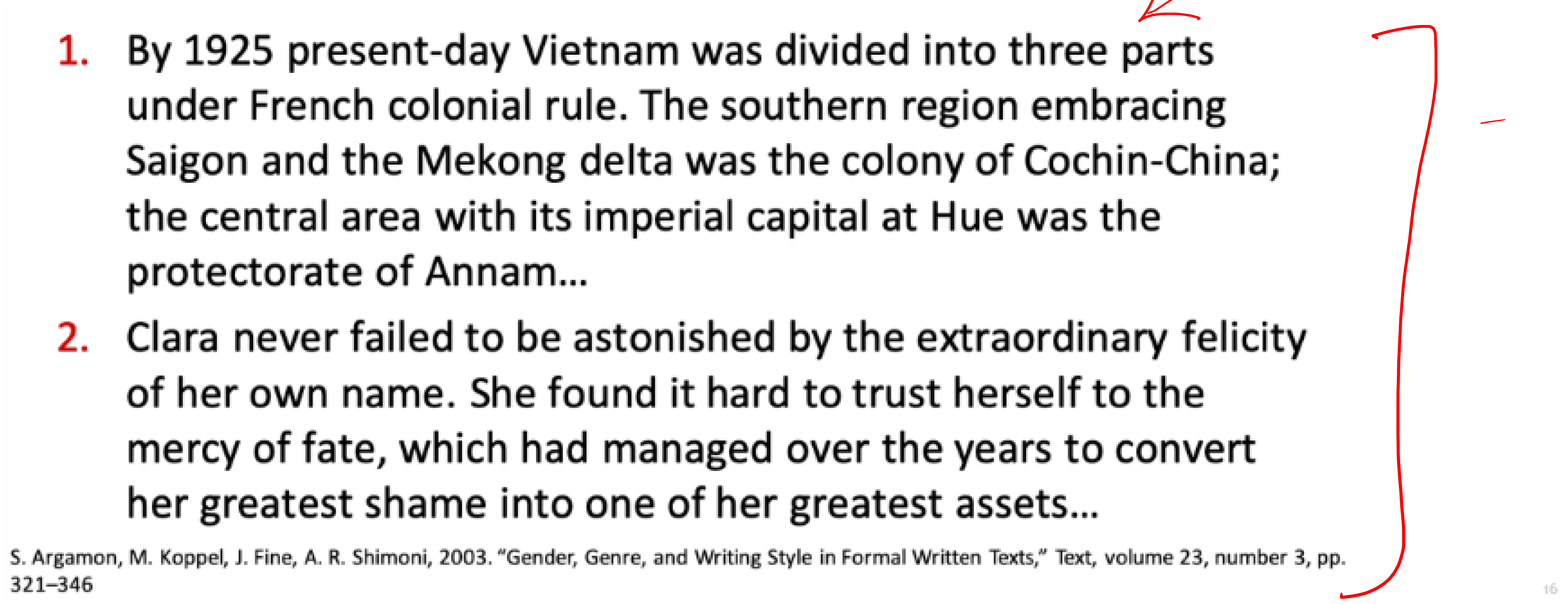
Andrea Chalupa @AndreaChalupa · 7m

Trump is betting on Iran doing something so horrific to Americans that we rally around the flag, and the 2020 election becomes a mindless debate of who's "patriotic" vs. who's anti-war ("weak" on Iran).

47 147 425

Female or Male author?

Author 1
2
3
.
9
Author 10

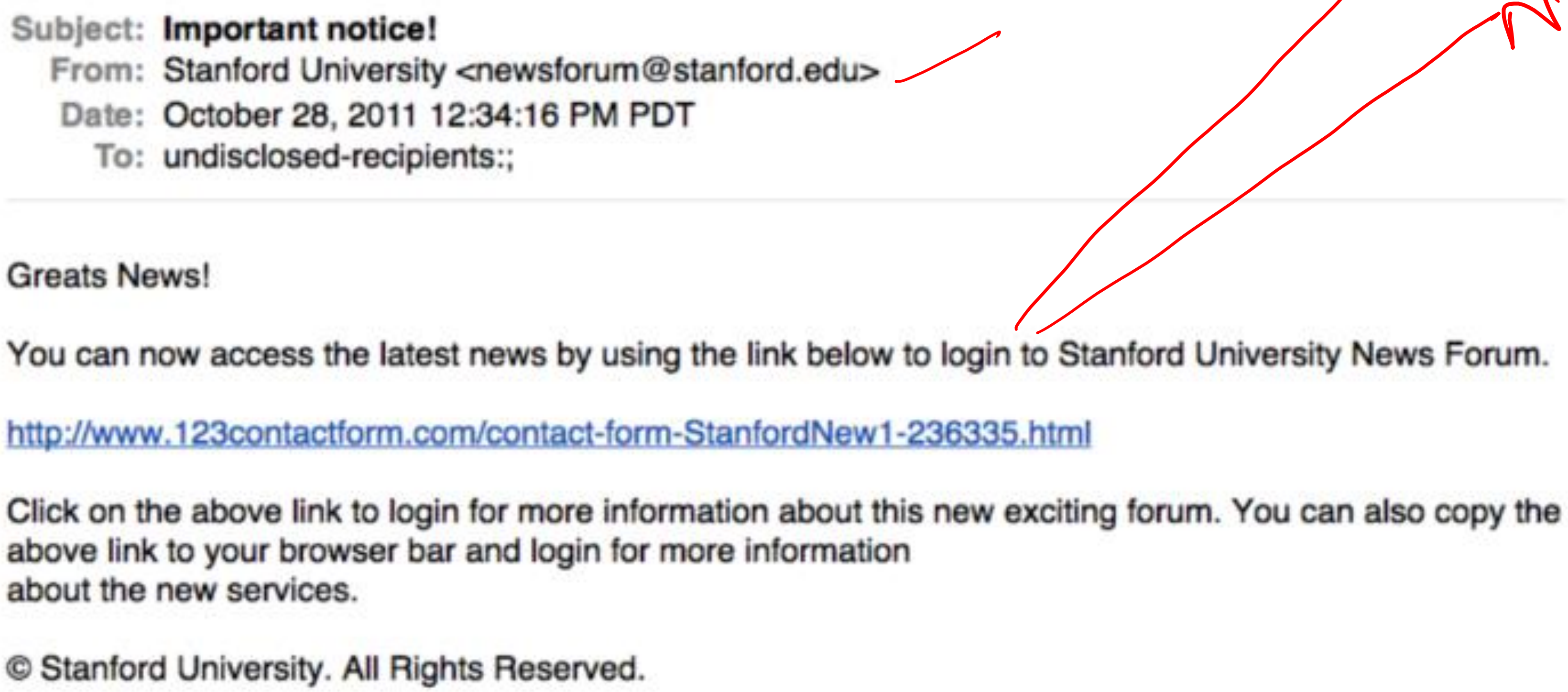


1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 321–346

16

Is this Spam?



Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

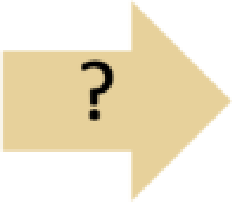
© Stanford University. All Rights Reserved.

Spam

No spam

What is the subject of this article?

MEDLINE Article



MeSH Subject Category Hierarchy

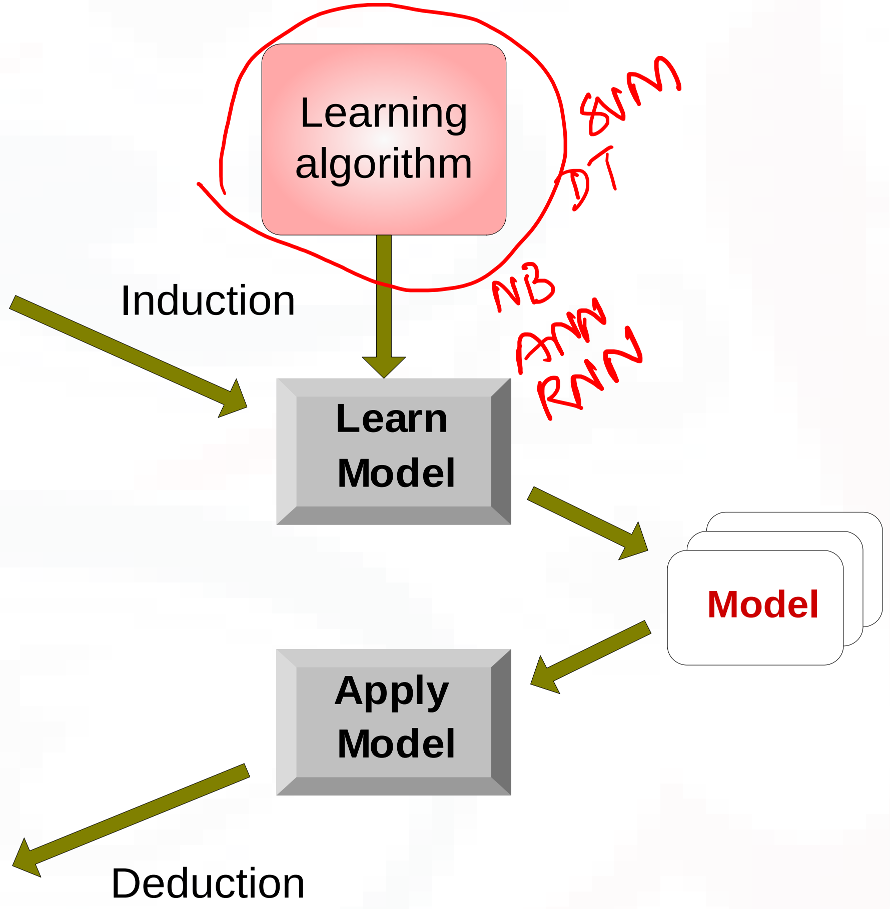
- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

18

Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

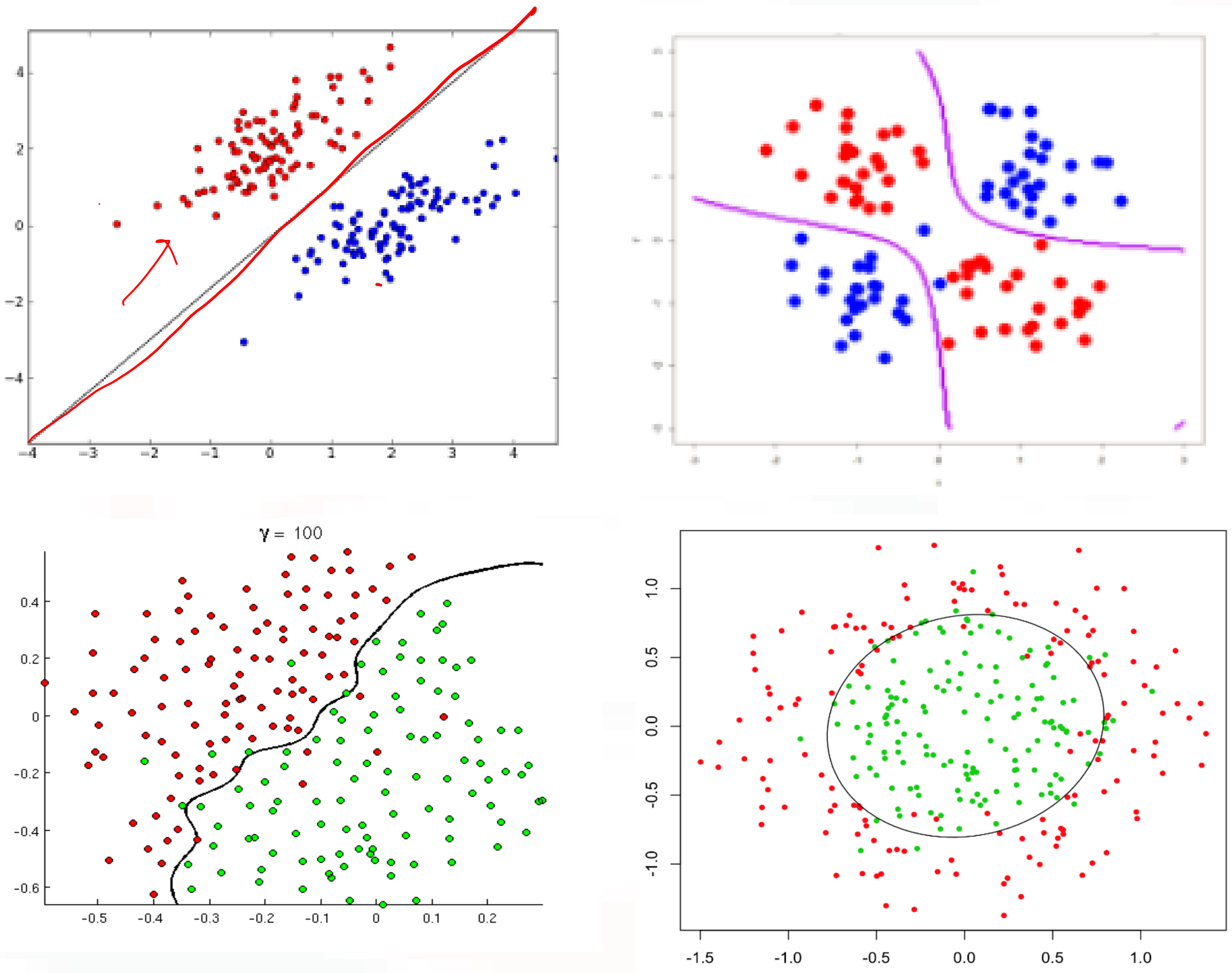


Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Tan, Steinbach, Kumar Introduction to Data Mining

Decision Boundaries



Converting text into Vector

Text Vectorization

- Bag of Words
- N-grams
- TF-IDF

Bag of Words (BOW)

Article — Document

Corpus —

- The Bag of Words (BoW) model is a text representation technique in which a document is represented as an unordered set, or "bag," of its words.
- The model involves:
  - creating a vocabulary of unique words from a corpus and
  - representing each document as a vector, with each dimension corresponding to a word in the vocabulary and
  - the value in each dimension representing the frequency of that word in the document.

walk

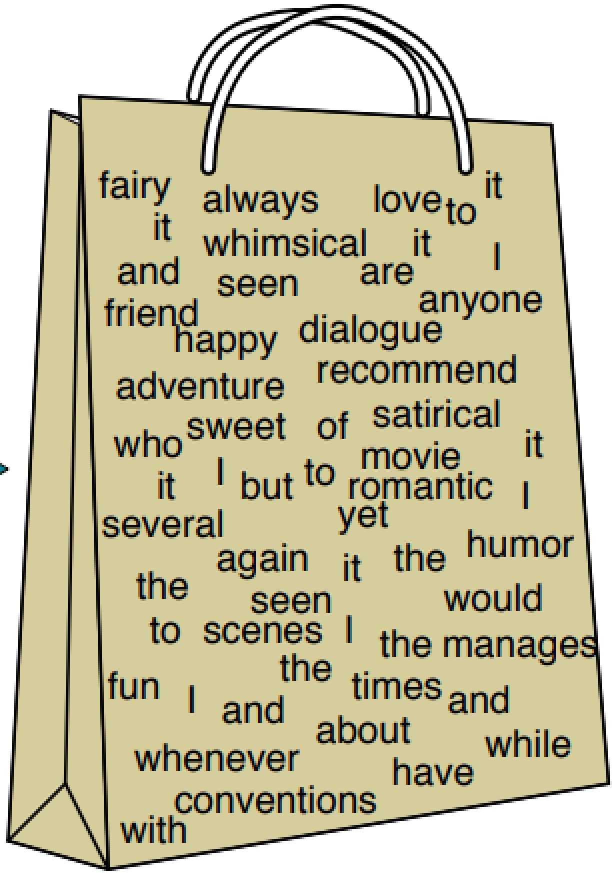
walked

walking /

walk

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



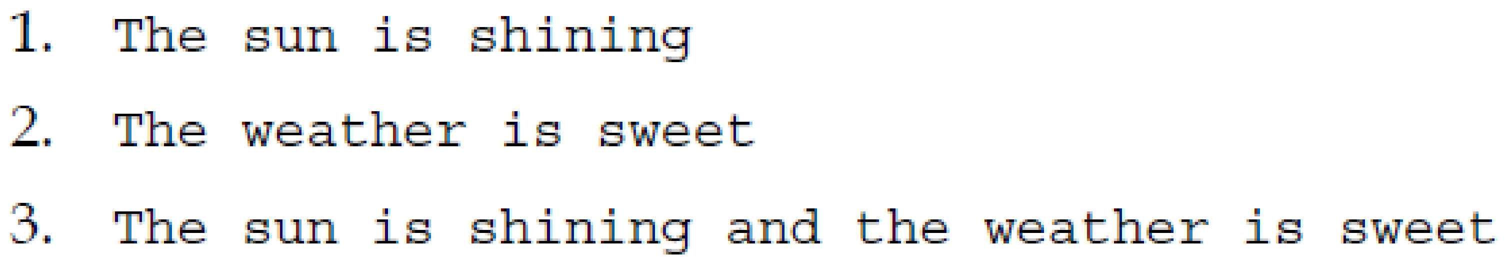
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Bag of words

Doc1

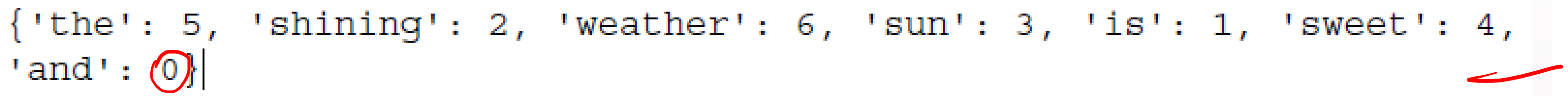
Doc2

Doc3



1. The sun is shining
2. The weather is sweet
3. The sun is shining and the weather is sweet

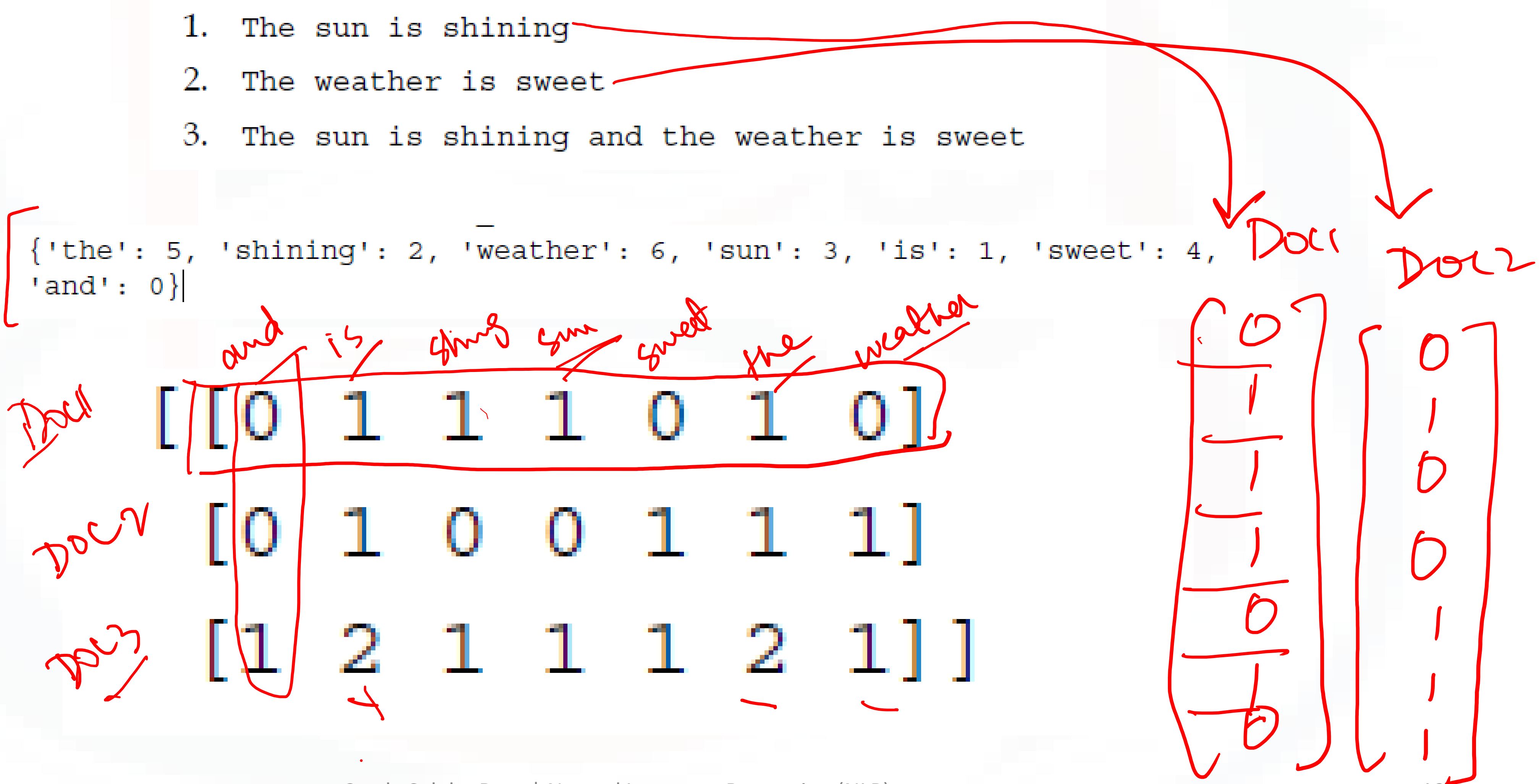
Corpus



```
{'the': 5, 'shining': 2, 'weather': 6, 'sun': 3, 'is': 1, 'sweet': 4,
'and': 0}
```

Vocabulary

Bag of Words (BOW)



1. The sun is shining
2. The weather is sweet
3. The sun is shining and the weather is sweet

```
{'the': 5, 'shining': 2, 'weather': 6, 'sun': 3, 'is': 1, 'sweet': 4,
'and': 0}
```

```
[[0 1 1 1 0 1 0]
 [0 1 0 0 1 1 1]
 [1 2 1 1 1 2 1]]
```

Weaknesses

- absence of semantic meaning and context
- some words are not weighted accordingly (“weather” weights less than the word “is” or “the”).

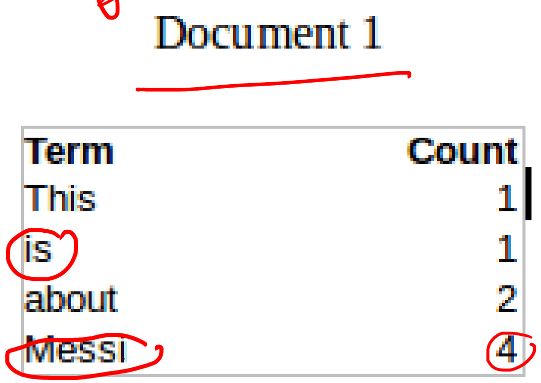
TF-IDF Score

- TF-IDF takes into account not just the occurrence of a word in a single document but in the entire corpus.

TF-IDF

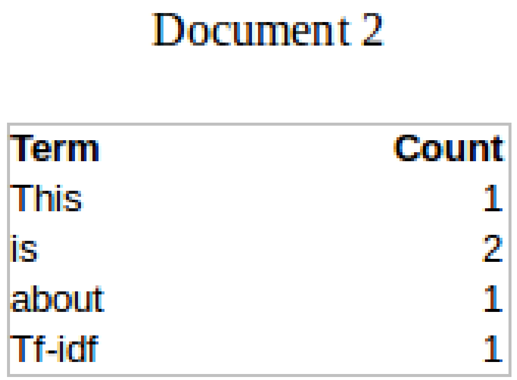
good

of the



Document 1

Term	Count
This	1
is	1
about	2
Messi	4



Document 2

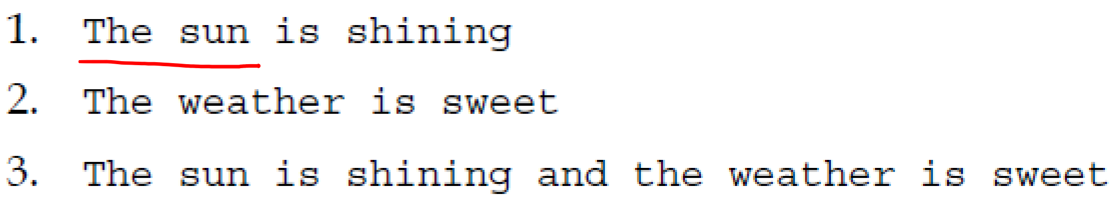
Term	Count
This	1
is	2
about	1
Tf-idf	1

- Term Frequency (TF) =
  - (Number of times term t appears in a document)/(Number of terms in the document)
- Inverse Document Frequency IDF =
  - log(N/n),

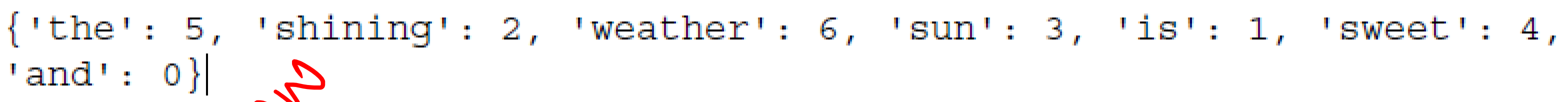
  where, N is the number of documents and n is the number of documents a term t has appeared in
- TF-IDF(This,Document1) = (1/8) * (0) = 0
- TF-IDF(This, Document2) = (1/5) * (0) = 0
- TF-IDF(Messi, Document1) = (4/8)*0.301 = 0.15

Saleness

TF-IDF

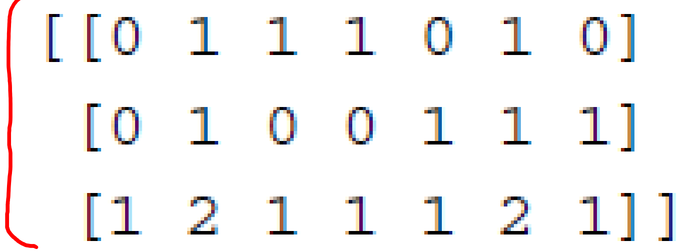


1. The sun is shining
2. The weather is sweet
3. The sun is shining and the weather is sweet



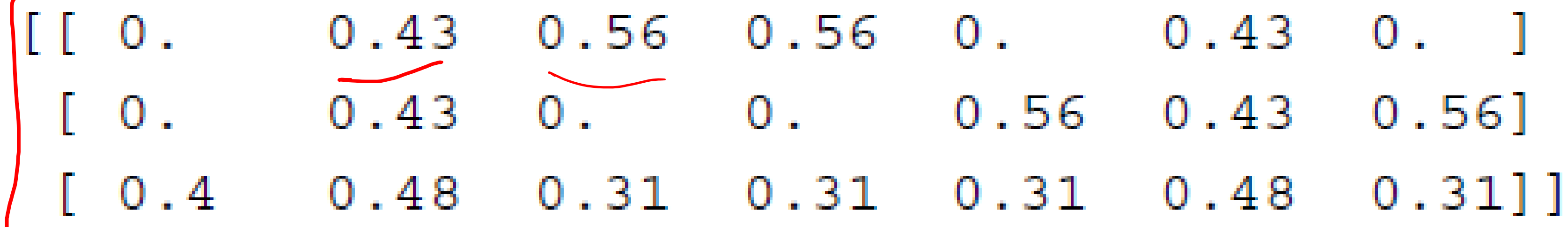
```
{'the': 5, 'shining': 2, 'weather': 6, 'sun': 3, 'is': 1, 'sweet': 4,
'and': 0}
```

BoW



```
[[0 1 1 1 0 1 0]
 [0 1 0 0 1 1 1]
 [1 2 1 1 1 2 1]]
```

Tf-IDf



```
[[ 0.    0.43  0.56  0.56  0.    0.43  0.  ]
 [ 0.    0.43  0.    0.    0.56  0.43  0.56]
 [ 0.4   0.48  0.31  0.31  0.31  0.48  0.31]]
```

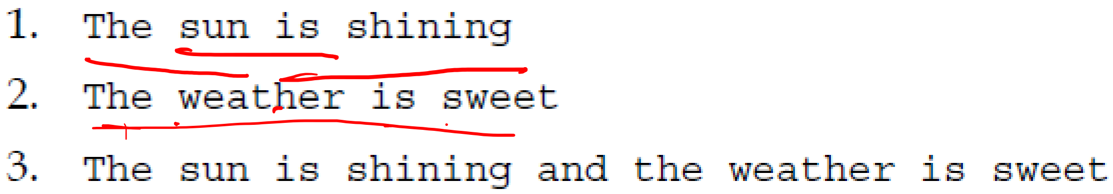
N-gram Model

2
3

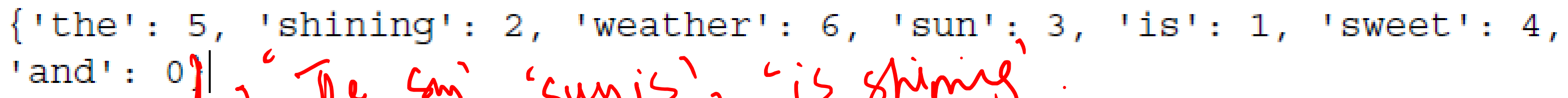
- An N-gram means a sequence of N words. Sometimes we want to consider the sequence of few words as a token:
  - Artificial Intelligence
  - Machine Learning
  - Cyber Security
  - Information retrieval
  - Natural Language Processing
  - South Africa
  - United Arab Emirates

Bag of Words (BOW)

South Africa



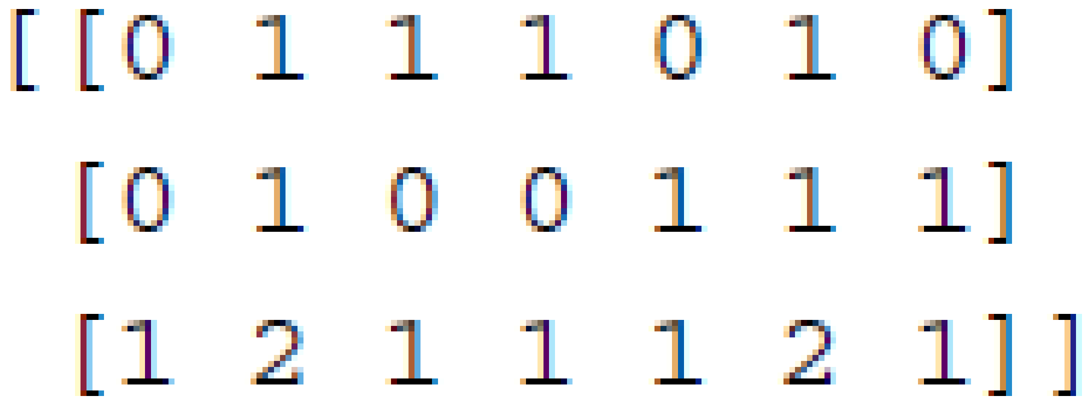
1. The sun is shining
2. The weather is sweet
3. The sun is shining and the weather is sweet



```
{'the': 5, 'shining': 2, 'weather': 6, 'sun': 3, 'is': 1, 'sweet': 4,
'and': 0}
```

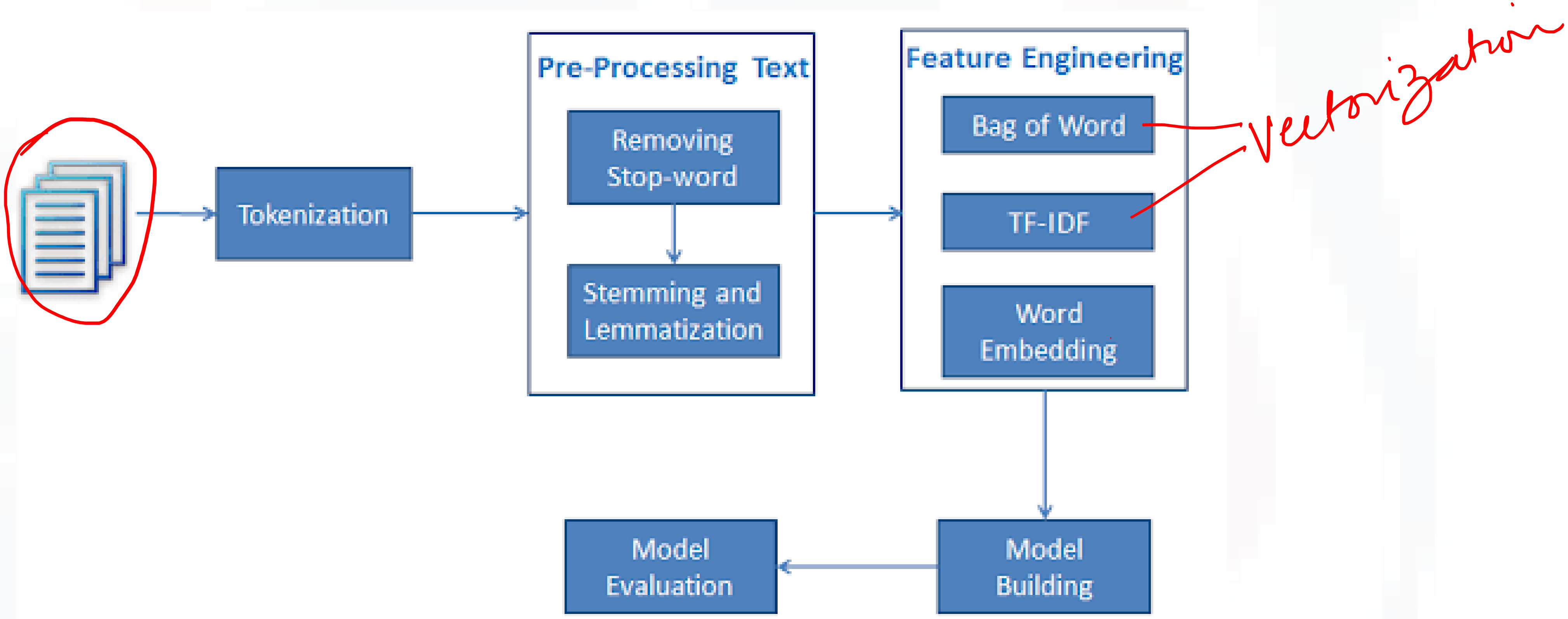
n=2

, 'The sun', 'sun is', 'is shining'.



```
[[0 1 1 1 0 1 0]
 [0 1 0 0 1 1 1]
 [1 2 1 1 1 2 1]]
```

Text Analysis



https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

References

- CS-4650/7650: Natural Language Processing (gatech.edu)
- http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
- https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

Thanks