

DSIF-4 Project 3

Project 3: Web APIs and NLP

...

By: Kwok Wei Hao

Presentation Outline

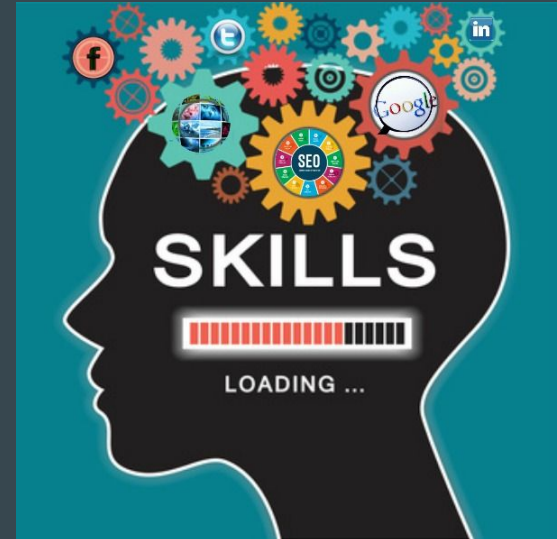
1. Background Information
2. Problem Statement
3. Data Gathering
4. Data Cleaning
5. Exploratory Data Analysis
6. Modelling
7. Results and Conclusion

Background Information

Rapid adoption of digital technologies, boosted by the COVID-19 crisis

Shift in skillsets required in the job market

Surging demand for data science and digital marketing to cope with the new normal.



Problem Statement

Business analyst with General Assembly

Identify if data science or digital marketing are the most talked about skillset

Provide instructional team with better understanding of the relevance/popularity

Better resource allocation to support either curriculum



Data Gathering



DS

Data Science

r/datascience

Scraped 20,000 posts per subreddit
with the use of Pushshift's API



Digital Marketing

r/DigitalMarketing

Data Cleaning

```
In [4]: 1 ds_df.shape
```

```
Out[4]: (19692, 86)
```

Keep just the
relevant features!



	title	selftext	subreddit
0	Hello everyone, what are the different methods and algorithms for model aggregation in federated learning? Can anyone help with any resource/articles...?	NaN	datascience
1	Is it worth starting Data Science as an Individual Contributor?	For context, I've worked with data for 10 years but mostly in analysis/reporting roles. I've recently moved into a people manager position at a Fortune 500 company in data management. In parallel,...	datascience
2	Imputing features like ratings and rankings	Can any one tell how to deal with null values for rankings and ratings features in a movie revenue dataset. \nI was thinking of imputing with mean values for groupings based on Genre but that does...	datascience
3	Best way to deal with missing/empty data in a small dataset	Hi. Potentially a simple, recurring questions here..\n\nI have a small dataset with around 10k rows. It has several columns, two of which have around 1.4k null values (no common null values between...	datascience
4	Curious how many of us work with data streaming or data batch	What are your thoughts on those two?\n\n[View Poll](https://www.reddit.com/poll/t5ose6)	datascience

Data Cleaning

selftext	subreddit
NaN	datascience
mostly in analysis/reporting roles. I've on at a Fortune 500 company in data management. In parallel,...	datascience
kings and ratings features in a movie values for groupings based on Genre but that does...	datascience

Replace NaN with empty string

title	selftext	subreddit
Associate (DEA-7TT2) Certification	[removed]	datascience
ow can I become a data scientist?	[removed]	datascience
ship + expected Entry-Level salary	[removed]	datascience
a entry, copy-paste, web research	[removed]	datascience
erve machine learning applications	[removed]	datascience
learn python	[removed]	datascience

Replace removed selftext with
empty string

title	selftext	subreddit
spent cleaning company database	[deleted]	datascience
test	[deleted]	datascience
New data science grad	[deleted]	datascience
unities for Aspiring Data Scientists	[deleted]	datascience
Some stats	[deleted]	datascience
utility of a second Master's Degree	[deleted]	datascience
Scientists come in many flavours.	[deleted]	datascience

Replace deleted selftext with
empty string

Data Cleaning

selftext
In this article, we talk about the application of CI/CD in machine ai/what-is-ci-cd-in-machine-learning/](https://learn.layer.ai/what-is-ci- cd-in-machine-learning/)
nd interested in data science for many years. but i don't have any job i need to join a team. but how?
ta science, and I am trying to learn data science from Harvard edx. In sked us to:\n\n" Create two five row vectors showing the 10th, 30th...
ry data?\n\nI have a some time series data about a battery and would like to know how to calculate the round trip efficiency (RTE).
ta science worth the time and money? Does it become a hindrance to

Remove URL links using RegEx

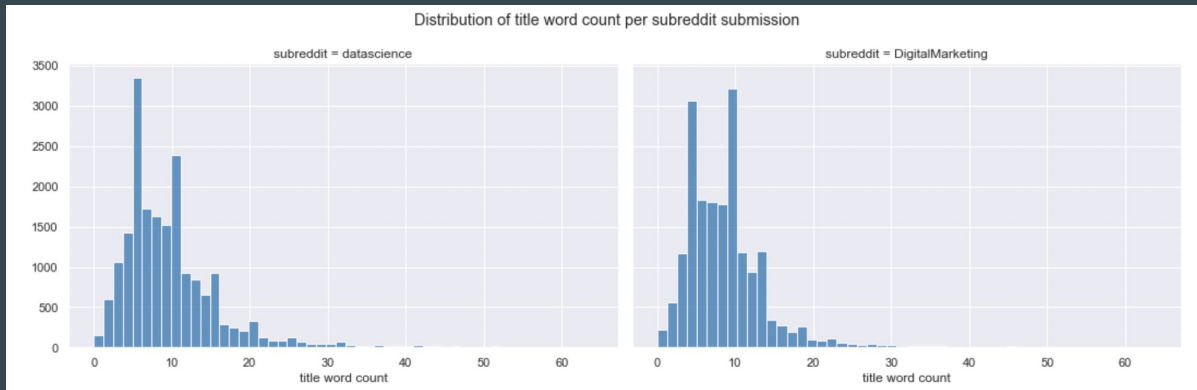
selftext
ry data?\n\nI have a some time series data about a battery and would like to know how to calculate the round trip efficiency (RTE).
a science worth the time and money? Does it become a hindrance to t have masters degree as a data scientist?\n\nI am a data scientist ...
problem from work and just couldn't find anything online for hours.\nI OST when I peaked at the sub's rules again and read with a booming voice in...
problem from work and just couldn't find anything online for hours.\nI POST about my problem when I peaked at the sub's rules again and read with a...

Remove newline characters

selftext	subreddit
a small dataset with around 10k rows. It has several 4k null values (no common null values between bo...	datascience
What are your thoughts on those two?[View Poll](datascience
[datascience
ow.Actual table starts from 4th row. I want to have a ant take away for Noodles.Horizontally, I can have ...	datascience
d positively and agreed to chat on a phone call.What hen & how should i ask for a referral /interview?	datascience
ree with a specialisation in Data Science. I know the	datascience

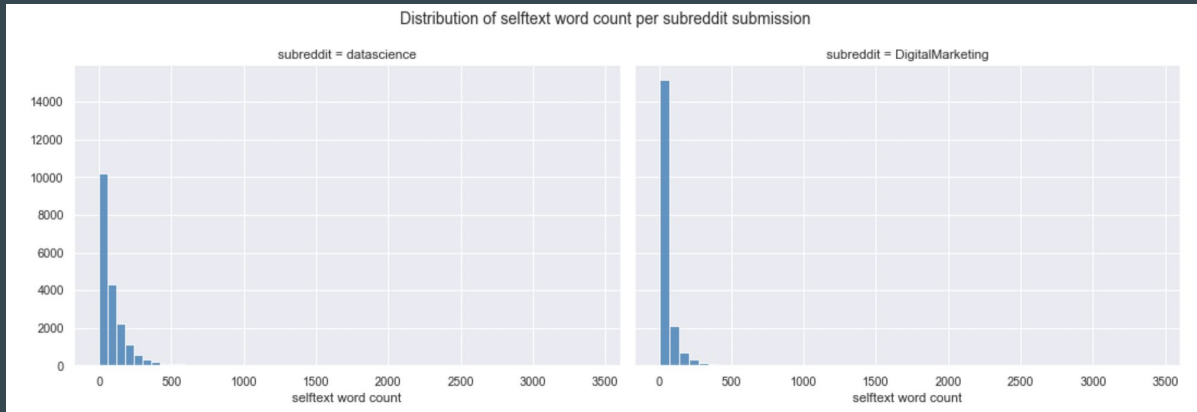
Remove '[View Poll]' characters

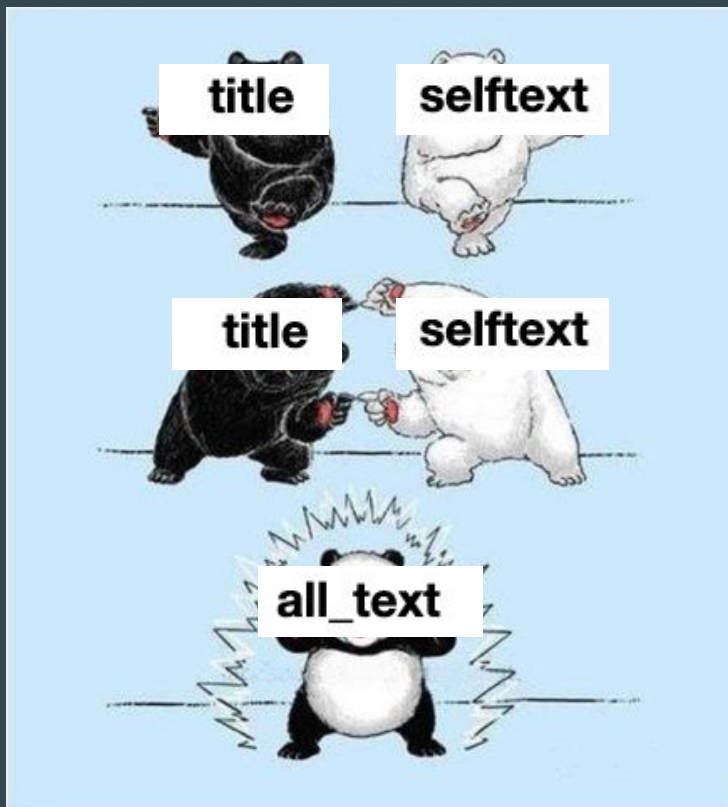
Exploratory Data Analysis



Right-skewed distribution

Outliers kept as they
should not affect model
adversely





Combine all text
data into 1 single
feature column!

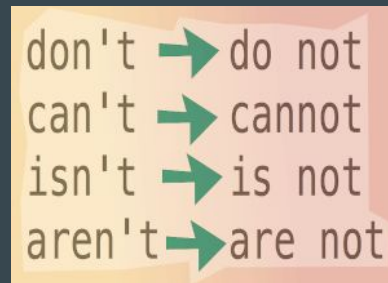
	all_text	subreddit
an anyone help with any resource/articles...?		datascience
ostly in analysis/reporting roles. I've recently moved into a people manager po...		datascience
es in a movie revenue dataset. I was thinking of imputing with mean values ...		datascience
a small dataset with around 10k rows. It has several columns, two of which ha...		datascience
atch What are your thoughts on those two?(datascience

Modelling - Preprocessing



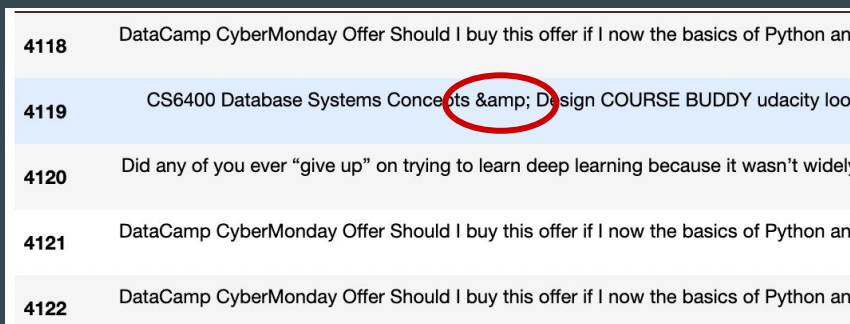
A blue rectangular box containing the text "HeLlO Python" at the top and "hello python" at the bottom. A red arrow points downwards from the first word of the top line to the first word of the bottom line, illustrating the conversion of text to lowercase.

Convert all text to lowercase



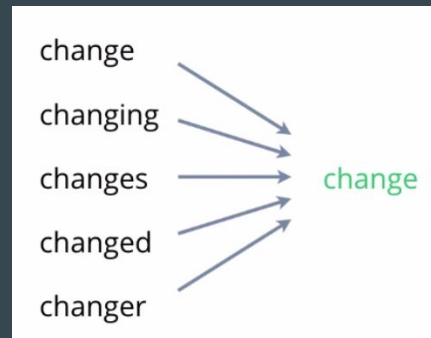
A light orange rectangular box containing four lines of text. Each line shows a contraction on the left, a green arrow pointing right, and the expanded form on the right: "don't" to "do not", "can't" to "cannot", "isn't" to "is not", and "aren't" to "are not".

Expand contracted words



A screenshot of a list of items, each with an ID and a text snippet. The text snippets are repeated. In the second item (ID 4119), the text "Concepts & Design" is visible, with the ampersand "&" circled in red to highlight it as a special character to be removed.

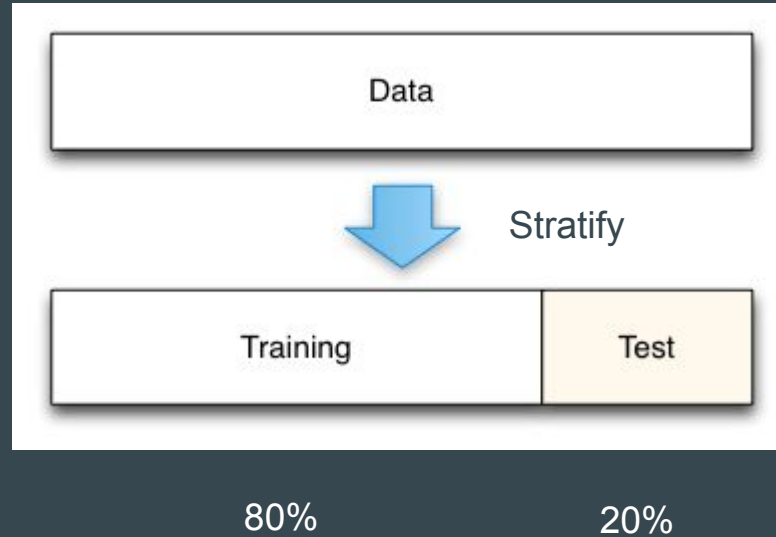
Remove all special characters and numbers with RegEx



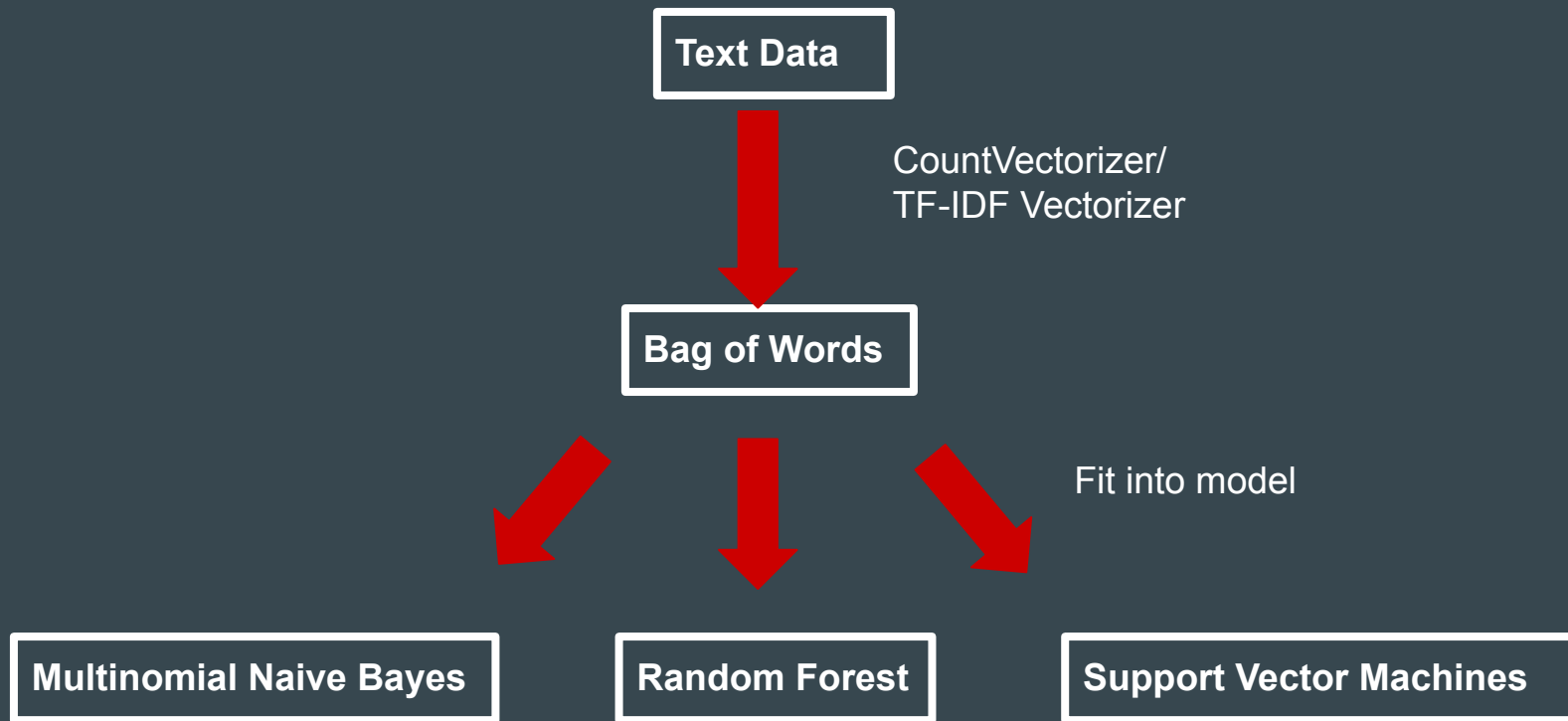
A white rectangular box showing the process of lemmatization. On the left, five words are listed: "change", "changing", "changes", "changed", and "changer". Blue arrows point from each of these words to a single word on the right, "change", which is colored green. This illustrates how different forms of a word are reduced to their base form.

Lemmatizing text data

Modelling - Train-test-split




Modelling



Results

Model	Train Accuracy Score	Test Accuracy Score
Baseline model	50.71%	
Multinomial Naive Bayes and CountVectorizer	94.75%	94.90%
Multinomial Naive Bayes and TF-IDF Vectorizer	94.63%	94.72%
Random Forest and CountVectorizer	91.40%	91.10%
Random Forest and TF-IDF Vectorizer	91.21%	90.69%
Support Vector Machines and CountVectorizer	96.14%	94.42%
Support Vector Machines and TF-IDF Vectorizer	98.55%	95.23%



95.23% test
accuracy

Conclusion

Model able to reliably classify text data belonging to either r/datascience or r/digital marketing with an accuracy of 95.23%, saves time for instructional team

However, being able to classify text data is only one part of the puzzle.

Suggested further improvements:

- Collect posts/comments over a 3-month period and establish post traffic for each subreddit
- Conduct sentiment analysis to analyze forummers' perspective/impression towards either topic