Table of Contents

In [ ]:
```
%%capture
%pip install pandas
%pip install sqlite3
```

Preparations

In [ ]:
```
import pandas as pd
```

Read data from tsv source

In [ ]:
```
df = pd.read_csv("./data/rspct.tsv", sep='\t', nrows=5000)
```

Check if data is correct

In [ ]:
```
df
```

Out[ ]:

	id	subreddit	title	selftext
0	6d8knd	talesfromtechsupport	Remember your command line switches...	Hi there, <lb>The usual. Long time lerker, fi...
1	58mbft	teenmom	So what was Matt "addicted" to?	Did he ever say what his addiction was or is h...
2	8f73s7	Harley	No Club Colors	Funny story. I went to college in Las Vegas. T...
3	6ti6re	ringdoorbell	Not door bell, but floodlight mount height.	I know this is a sub for the 'Ring Doorbell' b...
4	77sxto	intel	Worried about my 8700k small fft/data stress r...	Prime95 (regardless of version) and OCCT both,...
...	...	...	...	...
4995	5rg7ag	tdi	Buyback appointment was this morning. Picked u...	I turned in my '13 JSW this morning and she wi...
4996	6w4vz5	ableton	Where to get an Ableton Live Lite Key	Does anyone know a place where I can get an Ab...
4997	7fhc3t	TheLastAirbender	[All Spoilers] I just finished rewatching the ...	It's been bugging me since I started rewatchin...
4998	6efjh6	Dentistry	When to get a second opinion? New dentist want...	Hi Dentist of Reddit, <lb><lb>Is there any re...
4999	4pnb8m	unitedkingdom	Curious, why/how did West Midlands manage to g...	[Here] (http://www.bbc.co.uk/news/uk-politics-e...

5000 rows × 4 columns

Connect to database

In [ ]:
```
import sqlite3
conn = sqlite3.connect('/home/deepcode/Downloads/selfposts.db', timeout=10)
```

Save subreddit category info

In [ ]:
```
df_sr = pd.read_csv("./data/subreddit_info.csv")
```

In [ ]:
```
df_sr.to_sql("categories", con=conn, if_exists="replace", index=False)
```

Out[ ]: 3394

Cleaning process

Cleaning function

In [ ]:
```
import re
```

```
def clean(s):
    s = s.replace("<lb>", "\n")
    s = s.replace("\n", "")
    s = s.replace("<tab>", "\t")
    s = re.sub(r'<br */*>', "\n", s)
    s = s.replace("&lt;", "<").replace("&gt;", ">").replace("&amp;", "&")
    s = s.replace("&amp;", "&")
    # markdown urls
    s = re.sub(r'\(https*://[^\)]*\)', "", s)
    # normal urls
    s = re.sub(r'https*://[^\s]*', "", s)
    s = re.sub(r'_+', ' ', s)
    s = re.sub(r'"+', '"', s)
    return str(s)
```

Create new column in dataframe

In [ ]:
```
df["selftext_clean"] = ''
```

Iterate and clean

In [ ]:
```
for i, row in df.iterrows():
    df.at[i, "selftext_clean"] = clean(row.selftext)
```

Check results

In [ ]:
```
df.head()
```

Out[ ]:

	id	subreddit	title	selftext	selftext_clean	selftext_lemma	selftext_nc
0	6d8knd	talesfromtechsupport	Remember your command line switches...	Hi there, <lb>The usual. Long time lerker, fi...	Hi there, The usual. Long time lerker, first ...	hi there , \n the usual . long time lerker , ...	
1	58mbft	teenmom	So what was Matt "addicted" to?	Did he ever say what his addiction was or is h...	Did he ever say what his addiction was or is h...	did he ever say what his addiction was or is h...	
2	8f73s7	Harley	No Club Colors	Funny story. I went to college in Las Vegas. T...	Funny story. I went to college in Las Vegas. T...	funny story . i went to college in las vegas	
3	6ti6re	ringdoorbell	Not door bell, but floodlight mount height.	I know this is a sub for the 'Ring Doorbell' b...	I know this is a sub for the 'Ring Doorbell' b...	i know this is a sub for the ' ring doorbell '...	
4	77sxto	intel	Worried about my 8700k small fft/data stress r...	Prime95 (regardless of version) and OCCT both,...	Prime95 (regardless of version) and OCCT both,...	prime95 (regardless of version) and occt bot...	

NLP

Load spaCy

In [ ]:
```
%%capture
!python3 -m spacy download en-core-web-sm
```

In [ ]:
```
import spacy
nlp = spacy.load("en_core_web_sm")
```

Iterate over all rows and perform NLP

In [ ]:
```
for i, row in df.iterrows():
    if i % 1000 == 0:
        print(i)
    if(row["selftext_clean"] and len(str(row["selftext_clean"])) < 1000000):
        doc = nlp(str(row["selftext_clean"]))
        adjectives = []
        nouns = []
        verbs = []
```

```
        lemmas = []

        for token in doc:
            lemmas.append(token.lemma_)
            if token.pos_ == "ADJ":
                adjectives.append(token.lemma_)
            if token.pos_ == "NOUN" or token.pos_ == "PROPN":
                nouns.append(token.lemma_)
            if token.pos_ == "VERB":
                verbs.append(token.lemma_)

        df.at[i, "selftext_lemma"] = " ".join(lemmas)
        df.at[i, "selftext_nouns"] = " ".join(nouns)
        df.at[i, "selftext_adjectives"] = " ".join(adjectives)
        df.at[i, "selftext_verbs"] = " ".join(verbs)
        df.at[i, "selftext_nav"] = " ".join(nouns+adjectives+verbs)
        df.at[i, "no_tokens"] = len(lemmas)
```

```
0
1000
2000
3000
4000
```

Check results

In [ ]:
```
df.head()
```

Out[ ]:

	id	subreddit	title	selftext	selftext_clean	selftext_lemma	selftext_nc
0	6d8knd	talesfromtechsupport	Remember your command line switches...	Hi there, <lb>The usual. Long time lerker, fi...	Hi there, The usual. Long time lerker, first ...	hi there , the usual . long time lerker , fi...	time lerker poster p story devel
1	58mbft	teenmom	So what was Matt "addicted" to?	Did he ever say what his addiction was or is h...	Did he ever say what his addiction was or is h...	do he ever say what his addiction be or be he ...	addiction addict a group NA dr
2	8f73s7	Harley	No Club Colors	Funny story. I went to college in Las Vegas. T...	Funny story. I went to college in Las Vegas. T...	funny story . I go to college in Las Vegas . t...	story co Las V college b strip
3	6ti6re	ringdoorbell	Not door bell, but floodlight mount height.	I know this is a sub for the 'Ring Doorbell' b...	I know this is a sub for the 'Ring Doorbell' b...	I know this be a sub for the ' Ring Doorbell '...	sub Doc Floodlight bracket f
4	77sxto	intel	Worried about my 8700k small fft/data stress r...	Prime95 (regardless of version) and OCCT both,...	Prime95 (regardless of version) and OCCT both,...	Prime95 (regardless of version) and OCCT bot...	Prin version O test part b temp s

Save to database

In [ ]:
```
df.to_sql('posts_nlp', conn, if_exists="replace")
```

Out[ ]:
```
5000
```