HOMEWORK 8

NOELLE BROWN

MSDS 7337 SECTION 402

RETRIEVING REVIEWS – CODE FROM HW5

- **100 reviews scraped – about half positive, half negative reviews**
- **From IMDB – animated genre**
  - https://www.imdb.com/search/title?genres=animation&explore=title_type,genres&pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=fd0c0dd4-de47-4168-baa8-239e02fd9ee7&pf_rd_r=EJ85SEHS2WAZPGB6A52H&pf_rd_s=center-4&pf_rd_t=15051&pf_rd_i=genre&title_type=movie&sort=num_votes,desc&ref_=adv_explore_rhs
  - https://www.imdb.com/search/title?title_type=feature&user_rating=1.0,5.5&num_votes=25,&genres=animation&countries=us&sort=num_votes,desc

RETRIEVING REVIEWS – NORMALIZING

- **Reviews are normalized by expanding contractions, removing punctuation, and removing stop words**

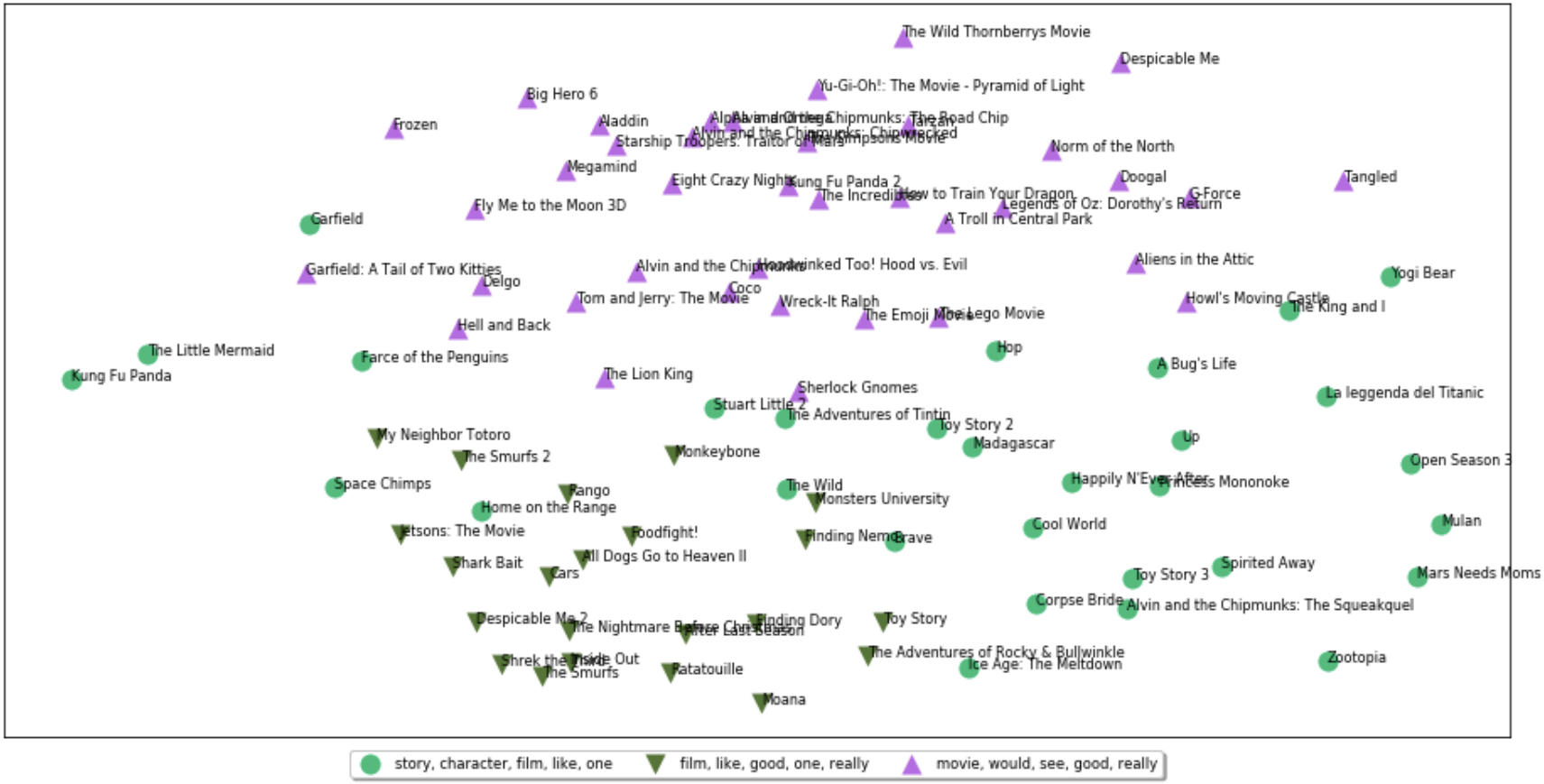
QUESTION 1 – SENTIMENT ANALYSIS

- SentiWordNet Sentiment Lexicon used
- Analysis retrieves positive, negative, and overall sentiment score between -1 (most negative) and 1 (most positive)
- A score of 0 indicates the sentiment is neutral
- This lexicon did a relatively good job of identifying words used in the reviews so I did not feel as though it was necessary to add words to this lexicon
- **Example output:**

```
------------------------------------------------------------
Review:
[u'toy', u'story', u'is', u'a', u'sheer', u'delight', u'to', u'view', u'on', u'the', u'screen', u'the', u'character
s', u'are', u'well', u'done', u'the', u'plot', u'is', u'exceptional', u'and', u'the', u'best', u'thing', u'of', u'a
ll', u'the', u'film', u'is', u'entirely', u'produced', u'on', u'the', u'computer', u'the', u'animation', u'is', u'e
xtraordinary', u'in', u'it', u'is', u'ability', u'to', u'bring', u'such', u'great', u'entertainment', u'to', u'th
e', u'screen', u'the', u'film', u'also', u'teaches', u'some', u'good', u'lessons', u'for', u'the', u'kids', u'lik
e', u'friendship', u'mainly', u'between', u'woody', u'and', u'buzz', u'lightyear', u'spectacular', u'entertainmen
t', u'all', u'around', u'and', u'one', u'of', u'the', u'best', u'films', u'disney', u'has', u'come', u'with']
     SENTIMENT STATS:
  Predicted Sentiment Objectivity Positive Negative Overall
0            positive         0.8     0.17     0.03    0.14
0.14
------------------------------------------------------------
```

QUESTION 2 – CLUSTER SENTIMENT ANALYSIS

- **I chose to use K-Means Clustering with 3 clusters**
  - I felt as though 3 clusters provided the most distinct clusters with the least amount of crossover from HW 7



QUESTION 2 – CLUSTER 1

```
Cluster 0 details:
--------------------
Key features: [u'story', u'character', u'film', u'like', u'one']
Movies in this cluster:
Up, Toy Story 3, Spirited Away, Toy Story 2, Kung Fu Panda, Zootopia, Brave, Madagascar, Princess Mononoke, A Bug's L
ife, Ice Age: The Meltdown, Corpse Bride, Mulan, The Little Mermaid, The Adventures of Tintin, Garfield, Stuart Littl
e 2, Alvin and the Chipmunks: The Squeakquel, Hop, The Wild, Mars Needs Moms, Cool World, Yogi Bear, Home on the Rang
e, Space Chimps, Happily N'Ever After, Open Season 3, Farce of the Penguins, The King and I, La leggenda del Titanic
Sentiment Scores:
0.02, 0.03, 0.06, -0.0, 0.03, 0.04, 0.03, 0.04, 0.01, 0.06, 0.05, 0.09, -0.04, 0.07, 0.05, 0.06, -0.0, 0.01, 0.01, 0.
03, 0.09, 0.03, 0.07, 0.01, -0.01, 0.05, -0.01, 0.02, 0.02, 0.02
Minimum Sentiment Score: -0.04
Mean Sentiment Score: 0.03133333333333334
Median Sentiment Score: 0.03
High Sentiment Score: 0.09
========================================
```

- This cluster has a sentiment range between -0.04 and 0.09, so overall tends to be relatively neutral
- The mean and the median score are both close to 0 at around 0.03 each
- The overall neutral sentiment of this cluster may explain the neutral words that describe this cluster – “story,” “character,” “film,” “like,” “one”
  - In a further analysis, I would consider removing words such as “film” since this does not add any value when considering movie reviews

QUESTION 2 – CLUSTER 2

```
Cluster 1 details:
--------------------
Key features: [u'film', u'like', u'good', u'one', u'really']
Movies in this cluster:
Finding Nemo, Toy Story, Ratatouille, Inside Out, Despicable Me 2, Cars, Monsters University, The Nightmare Before Ch
ristmas, Shrek the Third, My Neighbor Totoro, Rango, Moana, Finding Dory, The Smurfs, The Smurfs 2, The Adventures of
Rocky & Bullwinkle, Monkeybone, Foodfight!, After Last Season, All Dogs Go to Heaven II, Jetsons: The Movie, Shark Ba
it
Sentiment Scores:
0.02, 0.14, 0.04, 0.01, 0.02, 0.05, 0.03, 0.04, 0.0, 0.05, 0.03, 0.06, -0.01, 0.05, -0.02, 0.02, 0.04, -0.04, -0.02,
0.02, 0.05, 0.02
Minimum Sentiment Score: -0.04
Mean Sentiment Score: 0.02727272727272728
Median Sentiment Score: 0.025
High Sentiment Score: 0.14
========================================
```

- This cluster has a much larger range of sentiments with the lowest score of -0.04 and the highest score of 0.14
- The mean and the median scores also both fall around 0, at about 0.03 (rounded)
- The key features of this cluster do not seem to have much affect on or strong relation to the sentiments – “film,” “like,” “good,” “one,” “really
  - With key features of “like” and “good,” I would expect the mean and median sentiment scores to be more positive

QUESTION 2 – CLUSTER 3

```
Cluster 2 details:
--------------------
Key features: [u'movie', u'would', u'see', u'good', u'really']
Movies in this cluster:
The Lion King, How to Train Your Dragon, The Incredibles, Frozen, Despicable Me, Big Hero 6, Tangled, Wreck-It Ralph,
Aladdin, The Lego Movie, The Simpsons Movie, Howl's Moving Castle, Coco, Kung Fu Panda 2, Megamind, Alvin and the Chi
pmunks, The Emoji Movie, G-Force, Garfield: A Tail of Two Kitties, Alvin and the Chipmunks: Chipwrecked, Eight Crazy
Nights, Aliens in the Attic, Alvin and the Chipmunks: The Road Chip, The Wild Thornberrys Movie, Alpha and Omega, Tar
zan, Hoodwinked Too! Hood vs. Evil, Tom and Jerry: The Movie, Norm of the North, Yu-Gi-Oh!: The Movie - Pyramid of Li
ght, Sherlock Gnomes, A Troll in Central Park, Legends of Oz: Dorothy's Return, Doogal, Hell and Back, Starship Troop
ers: Traitor of Mars, Fly Me to the Moon 3D, Delgo
Sentiment Scores:
0.05, 0.12, 0.07, 0.05, 0.03, 0.06, 0.05, 0.03, 0.01, 0.05, 0.01, 0.03, 0.05, 0.02, 0.09, 0.03, 0.0, 0.04, 0.05, -0.0
2, 0.02, 0.01, 0.02, 0.09, 0.04, 0.01, 0.02, 0.02, 0.04, -0.02, 0.03, 0.01, 0.05, 0.03, 0.05, -0.04, 0.01, -0.03
Minimum Sentiment Score: -0.04
Mean Sentiment Score: 0.031052631578947373
Median Sentiment Score: 0.03
High Sentiment Score: 0.12
========================================
```

- This cluster also has a very wide range of scores, between -0.04 and 0.12, but the mean and the median remain similar
- Like the previous cluster, I would expect this cluster to have a higher positive mean and median sentiment due to the key features of “good” and “really.” The other features “movie,” “would,” and “see” are pretty neutral words – again, I would remove “movie” in further analyses
- It is interesting to note that the mean and median sentiment scores for all three clusters are similar (around 0.03). This may be evidence that the clusters do not take sentiment into account and are relatively random sentiment-wise

QUESTION 2 – OVERALL

The clusters do not seem to have distinct patterns based on the sentiment scores obtained from each. Each cluster has positive and negative scores along with similar mean and median scores. It appears as though the reviews are not clustering based on words that have strong sentiment meanings.

QUESTION 3A – SENTIMENT ANALYSIS ON CHUNKS

After chunking each of the reviews, I ran the chunks through my sentiment analyzer and saved the results as a table.

Chunk	Sentiment Score
[have, enjoyed]	0.31
[ranging]	0.13
[is]	0.13
[remains]	-0.13
[unparalleled]	0.38
[not]	-0.63
[creativity]	0.13
[are, equaled]	0.06
[truly]	0.63
[animated]	0.13
[is, usually, based]	0.04
[broad, slapstick]	0.19
[physical, exaggeration]	0.06
[are]	0.13
[their, beaks]	-0.06
[crowing]	0.38
[the, other, side]	-0.21
[such, sequences]	-0.06
[old, cartoon, conventions]	0.13
[sentient, animals]	-0.19

QUESTION 3B – SENTIMENT SCORES SORTED (HIGH TO LOW)

Most of the words that have the highest scores are positive. These words do a pretty good job of describing positive sentiment: "congratulations," "excellent," "happiness," "praise," "nice," "legendary," "better," "greatest and smartest," "happy." Several words are repeated, such as "better." It is interesting to note that most of these phrases are just adjectives while only a few of them are NP's or VP's.

Chunk	Sentiment Score
[Congratulations]	1
[the, excellent, Paul, Verhoeven]	1
[happiness]	1
[praise]	1
[nice, easy]	0.88
[important]	0.88
[the, legendary, Furious]	0.88
[legendary]	0.88
[better]	0.88
[better]	0.88
[greatest, and, smartest]	0.88
[better]	0.88
[better]	0.88
[better]	0.88
[happy]	0.88
[better]	0.88
[better]	0.88
[preferred]	0.88
[amused]	0.88
[better]	0.88

QUESTION 3B – SENTIMENT SCORES SORTED (LOW TO HIGH)

In the top 20 negative chunks, we have words such as "miserable," "not," "unfortunately," "negative," "fear," "awful," "atrociously awful," "crappy," "hard," and "stupid." The words that are repeated are "unfortunately," "hard," and "awful." These seem to do a pretty good job of describing negative chunks. Overall, it seems as though performing sentiment analysis on the POS chunks is more valuable than the sentiment analysis of the clusters.

Chunk	Sentiment Score
[miserable]	-0.88
[not, sure, I]	-0.88
[Unfortunately]	-0.88
[negative]	-0.88
[Unfortunately]	-0.88
[fear]	-0.88
[Unfortunately]	-0.88
[Unfortunately]	-0.88
[so, awful, it]	-0.88
[The, evil, controlling, it]	-0.88
[Unfortunately]	-0.88
[Unfortunately]	-0.88
[atrociously, awful]	-0.81
[crappy]	-0.75
[hard]	-0.75
[stupid, do, you]	-0.75
[their, hardest, I]	-0.75
[hard]	-0.75
[protecting]	-0.75
[dead]	-0.75

TOOLS USED

```
# Necessary imports
import platform; print platform.platform()
import sys; print "Python", sys.version
import nltk; print "nltk", nltk.__version__
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.corpus import sentiwordnet as swn
import bs4; print "BS4", bs4.__version__
from bs4 import BeautifulSoup, SoupStrainer
import requests; print "requests", requests.__version__
import urllib2; print "urllib2", urllib2.__version__
from urllib2 import Request, urlopen
import re; print "re", re.__version__
import os; print os.environ['CONDA_DEFAULT_ENV']
import numpy as np; print "numpy", np.__version__
import scipy; print "scipy", scipy.__version__
from scipy.stats import itemfreq
from scipy.cluster.hierarchy import ward, dendrogram
import copy
import pandas as pd; print "pandas", pd.__version__
import sklearn; print "sklearn", sklearn.__version__
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.manifold import MDS
from sklearn.metrics.pairwise import cosine_similarity
import string
import matplotlib.pyplot as plt
from matplotlib.font_manager import FontProperties
import random
import pattern; print "pattern", pattern.__version__
from pattern.en import parsetree
```

```
Darwin-17.5.0-x86_64-i386-64bit
Python 2.7.15 |Anaconda, Inc.| (default, Oct 23 2018, 13:35:16)
[GCC 4.2.1 Compatible Clang 4.0.1 (tags/RELEASE_401/final)]
nltk 3.3
BS4 4.6.3
requests 2.0.1
urllib2 2.7
re 2.2.1
Python2
numpy 1.15.3
scipy 1.1.0
pandas 0.23.4
sklearn 0.20.0
pattern 2.4
```