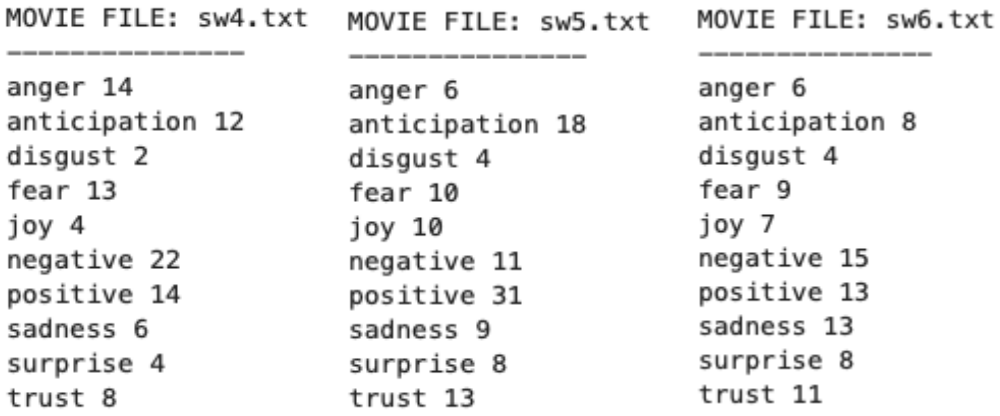
Kevin King
Professor Rolando Coto-Solano
COSC 72: Accelerated Computational Linguistics
23 April, 2023

Homework 3 Written Submission

Exercise 1: Darth Vader's Feelings

Based on the output results of my code, it was interesting to see the progression of Darth Vader's feelings throughout the original Star Wars trilogy. In sw4.txt, he exemplifies the most anger with a value of 14, whereas the other two scripts each had values of 6. Additionally, he gets sadder over time, going from a score of 6 to 9 to 13. Below are screenshots of the results for each text file:



```
MOVIE FILE: sw4.txt
---------------
anger 14
anticipation 12
disgust 2
fear 13
joy 4
negative 22
positive 14
sadness 6
surprise 4
trust 8
```

```
MOVIE FILE: sw5.txt
---------------
anger 6
anticipation 18
disgust 4
fear 10
joy 10
negative 11
positive 31
sadness 9
surprise 8
trust 13
```

```
MOVIE FILE: sw6.txt
---------------
anger 6
anticipation 8
disgust 4
fear 9
joy 7
negative 15
positive 13
sadness 13
surprise 8
trust 11
```

Exercise 2: Clustering Shakespeare's Plays

The clusters seem to somewhat make sense as similar plays are clustered with one another. For instance, five of the King Henry plays are grouped together in cluster 2, where the top term is "king," while the other two are grouped together in cluster 4. Additionally, the other plays in which the titles have the word "King" in them are in cluster 2. This would make sense because the plays with the word "King" in their names would likely have the word "king" in the script as well. This would explain why the new document 1 with the text "battle and king" is predicted to be in cluster 2. Document 2, which has the text "wit and love," is predicted to be in cluster 1, where the second top term is "love." Below are screenshots of my output results as well as the hierarchical clustering dendrogram:

Play Titles + Clusters

```
AllsWellThatEndsWell 2
AntonyCleopatra 3
AsYouLikeIt 1
ComedyErrors 1
Coriolanus 1
Cymbeline 1
Hamlet 5
KingHenry4.1 4
KingHenry4.2 4
KingHenry5 2
KingHenry6.1 2
KingHenry6.2 2
KingHenry6.3 2
KingHenry8 2
KingJohn 2
JuliusCaesar 3
KingLear 2
LovesLabourLost 1
MacBeth 8
MeasureForMeasure 2
MerchantVenice 0
WivesWindsor 4
MidsummerNightsDream 1
MuchAdo 6
Othello 1
Pericles 9
KingRichard2 2
KingRichard3 2
RomeoJuliet 1
TamingShrew 1
Tempest 1
Timon 1
TitusAndronicus 1
TroilusCressida 1
12Night 7
GentlemenVerona 1
NobleKinsmen 1
WintersTale 1
LoversComplaint 1
PassionatePilgrim 1
PhoenixTurtle 1
VenusAdonis 1
```

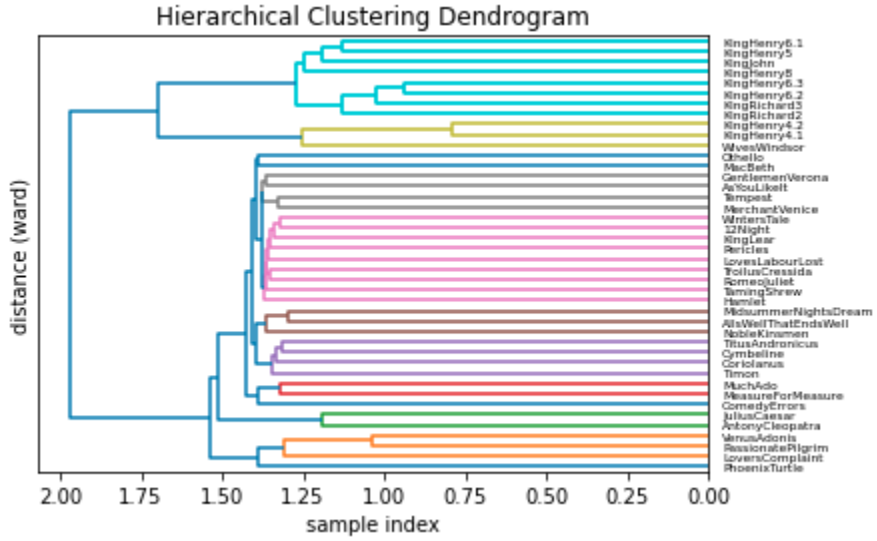
Top Terms Per Cluster

```
Top terms per cluster:
Cluster 0:
 portia
 bassanio
 shylock
 launcelot
 lorenzo
 antonio
 gratiano
 nerissa
 jessica
 salerio
Cluster 1:
 thou
 love
 thy
 thee
 shall
 did
 good
 like
 sir
 timon
Cluster 2:
 king
 thou
 gloucester
 thy
 lord
 henry
 shall
 york
 richard
 duke
Cluster 3:
 antony
 caesar
 brutus
 cassius
 cleopatra
 enobarbus
 charmian
 casca
 thou
 shall
Cluster 4:
 falstaff
 ford
 bardolph
 prince
 page
 thou
 hotspur
 mrs
 sir
 poins
Cluster 5:
 hamlet
 horatio
 polonius
 laertes
 ophelia
 rosencrantz
 guildenstern
 lord
 king
 marcellus
Cluster 6:
 benedick
 leonato
 beatrice
 pedro
 claudio
 don
 hero
 dogberry
 borachio
 margaret
Cluster 7:
 toby
 olivia
 viola
 malvolio
 sir
 aguecheek
 fabian
 maria
 clown
 sebastian
Cluster 8:
 macbeth
 macduff
 banquo
 malcolm
 ross
 duncan
 lennox
 murtherer
 thane
 lady
Cluster 9:
 pericles
 marina
 simonides
 helicanus
 thaisa
 boult
 lysimachus
 cleon
 cerimon
 fisherman
```

Predictions For New Documents

```
Predictions
Doc1 cluster: [2]
Doc2 cluster: [1]
```

Hierarchical Clustering Dendrogram



Exercise 3: Word Embeddings

I chose Spanish for the word embedding exercise. The difference between the embeddings for "hombre" (“man”) and "mujer" (“woman”) seems to be captured in the model's gender associations. It is likely that the model has learned that the word "hombre" is more strongly associated with masculinity and male-related concepts, while the word "mujer" is more strongly associated with femininity and female-related concepts. The operation "mujer + rey - hombre" ("woman+king-man") is intended to compute a vector that represents the concept of "reina" (“queen”), which is the first result given in the output (see next page). Thus, it seems like the arithmetic operation works as expected. Below are screenshots of my output results and the t-SNE chart:

Top 25 Most Similar Words to “Man” and “Woman” in Spanish

```
embeddings.get_nearest_neighbors('hombre', k=25)
```

```
[(0.6985235214233398, 'hombre.El'),
 (0.6847670078277588, 'mujer'),
 (0.6824022531509399, 'varón'),
 (0.6381804943084717, 'unhombre'),
 (0.6375911831855774, 'Hombre'),
 (0.6296564936637878, 'individuo'),
 (0.6243440508842468, 'humano'),
 (0.6197197437286377, 'elhombre'),
 (0.618654727935791, 'muchacho'),
 (0.6134461164474487, 'hombres'),
 (0.6023065447807312, 'hombre.Este'),
 (0.5991070866584778, 'no-hombre'),
 (0.5958799719810486, 'hombra'),
 (0.5950770974159241, 'hombre.Pero'),
 (0.5942564010620117, 'hombre.En'),
 (0.5892860293388367, 'hombre.Es'),
 (0.5885716676712036, 'niño-hombre'),
 (0.5856088399887085, 'hombre.La'),
 (0.584286093711853, 'hombre-'),
 (0.5841028690338135, 'mujer.El'),
 (0.58305823802948, 'hombre.Y'),
 (0.5821053981781006, 'chico'),
 (0.580610454082489, 'anciano'),
 (0.580307126045227, 'joven'),
 (0.5799315571784973, 'hombrees')]
```

```
embeddings.get_nearest_neighbors('mujer', k=25)
```

```
[(0.7110328078269958, 'mujer.La'),
 (0.6897695064544678, 'mujermujer'),
 (0.6861298084259033, 'mujer.Pero'),
 (0.684766948223114, 'hombre'),
 (0.6845912933349609, 'muchacha'),
 (0.6795122623443604, 'fémina'),
 (0.6703848242759705, 'lamujer'),
 (0.6586087942123413, 'esposa'),
 (0.6574076414108276, 'chica'),
 (0.6479310989379883, 'niña'),
 (0.6458329558372498, 'dama'),
 (0.6451638340950012, 'unamujer'),
 (0.6437935829162598, 'mujera'),
 (0.6435273885726929, 'Mujer'),
 (0.6399717926979065, 'mujer-mujer'),
 (0.6391507983207703, 'mujer.Es'),
 (0.6384367346763611, 'mujer.Esta'),
 (0.6358761787414551, 'varón'),
 (0.634837806224823, 'hija'),
 (0.6306506991386414, 'mujerde'),
 (0.6281776428222656, 'mujer.El'),
 (0.6268007755279541, 'persona'),
 (0.6255276203155518, 'mujer.Una'),
 (0.6250082850456238, 'mujer.En'),
 (0.6209526062011719, 'hombruna')]
```

Top 25 Results for "King - Man + Woman" in Spanish

```
embeddings.get_analogies("rey", "hombre", "mujer", k=25)
```

```
[(0.6996281743049622, 'reina'),
 (0.6584349870681763, 'princesa'),
 (0.578596293926239, 'reina-madre'),
 (0.5746439695358276, 'monarca'),
 (0.5572191476821899, 'emperatriz'),
 (0.5523837804794312, 'Rey'),
 (0.5444003939628601, 'reyes'),
 (0.5441058278083801, 'hija'),
 (0.5410926938056946, 'Reina'),
 (0.5355700254440308, 'consorte'),
 (0.5331939458847046, 'infanta'),
 (0.5261333584785461, 'reina-viuda'),
 (0.5260338187217712, 'esposa'),
 (0.5179920792579651, 'príncipe'),
 (0.5175434947013855, 'dama'),
 (0.517275333404541, 'infanta-reina'),
 (0.5155842304229736, 'emperadora'),
 (0.515200674533844, 'lareina'),
 (0.5045839548110962, 'laprincesa'),
 (0.504417359828949, 'virreina'),
 (0.5041970610618591, 'reyna'),
 (0.5037978887557983, 'realeza'),
 (0.502633273601532, 'monarquía'),
 (0.5008916258811951, 'reinona'),
 (0.4997826814651489, 'emperatríz')]
```

t-SNE Chart for [‘hombre’, ‘mujer’, ‘rey’, ‘reina’, ‘niño’, ‘chico’, ‘chica’]