Lab sheet 9 - Document Clustering

Kavishka Gamage – 17000475

Term Similarity

Hamming distance

Elephant term gives error when finding text similarity with believe

```
Features: ['a', 'b', 'e', 'g', 'h', 'i', 'l', 'n', 'p', 'r', 't', 'v']

root: [0 1 3 0 0 1 1 0 0 0 0 1]
term1: [0 1 3 0 0 1 1 0 0 0 0 1]
term2: [2 1 0 1 0 1 0 1 0 1 0 0]
term3: [1 0 2 0 1 0 1 1 1 0 1 0]

Hamming distance between root: Believe and term: beleive is 2
Hamming distance between root: Believe and term: bargain is 6
Traceback (most recent call last):
  File "g:/Github/NLP/UCSC/Labsheets/4.2. Document Clustering/4.2. Document Clustering/term_similarity.py", line 146, in <module>
    hamming_distance(root_vector, vector_term, norm=False)))
  File "g:/Github/NLP/UCSC/Labsheets/4.2. Document Clustering/4.2. Document Clustering/term_similarity.py", line 69, in hamming_distance
    raise ValueError('The vectors must have equal lengths.')
ValueError: The vectors must have equal lengths.
```

Manhattan distance

This measure also gave an error when terms don't have equal length

```
Features: ['a', 'b', 'e', 'g', 'h', 'i', 'l', 'n', 'p', 'r', 't', 'v']

root: [0 1 3 0 0 1 1 0 0 0 0 1]
term1: [0 1 3 0 0 1 1 0 0 0 0 1]
term2: [2 1 0 1 0 1 0 1 0 1 0 0]
term3: [1 0 2 0 1 0 1 1 1 0 1 0]

Manhattan distance between root: Believe and term: beleive is 8
Manhattan distance between root: Believe and term: bargain is 38
Traceback (most recent call last):
  File "g:/Github/NLP/UCSC/Labsheets/4.2. Document Clustering/4.2. Document Clustering/term_similarity.py", line 159, in <module>
    manhattan_distance(root_vector, vector_term, norm=False)))
  File "g:/Github/NLP/UCSC/Labsheets/4.2. Document Clustering/4.2. Document Clustering/term_similarity.py", line 74, in manhattan_distance
    raise ValueError('The vectors must have equal lengths.')
ValueError: The vectors must have equal lengths.
```

Euclidian distance

```
Features: ['a', 'b', 'e', 'g', 'h', 'i', 'l', 'n', 'p', 'r', 't', 'v']

root: [0 1 3 0 0 1 1 0 0 0 0 1]
term1: [0 1 3 0 0 1 1 0 0 0 0 1]
term2: [2 1 0 1 0 1 0 1 0 1 0 0]
term3: [1 0 2 0 1 0 1 1 1 0 1 0]

Euclidean distance between root: Believe and term: beleive is 5.66
Euclidean distance between root: Believe and term: bargain is 17.94
Traceback (most recent call last):
  File "g:/Github/NLP/UCSC/Labsheets/4.2. Document Clustering/4.2. Document Clustering/term_similarity.py", line 171, in <module>
    round(euclidean_distance(root_vector, vector_term),2)))
  File "g:/Github/NLP/UCSC/Labsheets/4.2. Document Clustering/4.2. Document Clustering/term_similarity.py", line 79, in euclidean_distance
    raise ValueError('The vectors must have equal lengths.')
ValueError: The vectors must have equal lengths.
```

Levenshtein edit distance

```
Computing distance between root: Believe and term: beleive
Levenshtein edit distance is 2
The complete edit distance matrix is depicted below
   b  e  l  i  e  v  e
b  0  1  2  3  4  5  6
e  1  0  1  2  3  4  5
l  2  1  0  1  2  3  4
e  3  2  1  1  1  2  3
i  4  3  2  1  2  2  3
v  5  4  3  2  2  2  3
e  6  5  4  3  2  3  2
----------------------------
Computing distance between root: Believe and term: bargain
Levenshtein edit distance is 6
The complete edit distance matrix is depicted below
   b  e  l  i  e  v  e
b  0  1  2  3  4  5  6
a  1  1  2  3  4  5  6
r  2  2  2  3  4  5  6
g  3  3  3  3  4  5  6
a  4  4  4  4  4  5  6
i  5  5  5  4  5  5  6
n  6  6  6  5  5  6  6
```

Cosine similarity

```
Analyzing similarity between root: Believe and term: beleive
Cosine distance  is -0.0
Cosine similarity  is 1.0
---------------------------------------
Analyzing similarity between root: Believe and term: bargain
Cosine distance  is 0.82
Cosine similarity  is 0.18000000000000005
---------------------------------------
Analyzing similarity between root: Believe and term: Elephant
Cosine distance  is 0.39
Cosine similarity  is 0.61
---------------------------------------
```

Document Similarity

Cosine similarity

```
Document Similarity Analysis using Cosine Similarity
==========================================================
Document 1 : The fox is definitely smarter than the dog
Top 2 similar docs:
----------------------------------------
Doc num: 8 Similarity Score: 1.0
Doc: The dog is smarter than the fox
----------------------------------------
Doc num: 7 Similarity Score: 0.426
Doc: The fox is quicker than the lazy dog
----------------------------------------

Document 2 : Java is a static typed programming language unlike Python
Top 2 similar docs:
----------------------------------------
Doc num: 5 Similarity Score: 0.733
Doc: Python and Java are popular Programming languages
----------------------------------------
Doc num: 6 Similarity Score: 0.58
Doc: Among Programming languages, both Python and Java are the most used in Analytics
----------------------------------------
```

Hellinger-Bhattacharya distance

```
Document Similarity Analysis using Hellinger-Bhattacharya distance
============================================================
Document 1 : The fox is definitely smarter than the dog
Top 2 similar docs:
----------------------------------------
Doc num: 8 Distance Score: 0.0
Doc: The dog is smarter than the fox
----------------------------------------
Doc num: 7 Distance Score: 0.96
Doc: The fox is quicker than the lazy dog
----------------------------------------

Document 2 : Java is a static typed programming language unlike Python
Top 2 similar docs:
----------------------------------------
Doc num: 5 Distance Score: 0.702
Doc: Python and Java are popular Programming languages
----------------------------------------
Doc num: 4 Distance Score: 0.925
Doc: Python is a great Programming language
----------------------------------------

Document 3 : I love to relax under the beautiful blue sky!
Top 2 similar docs:
----------------------------------------
Doc num: 2 Distance Score: 0.0
Doc: The sky is blue and beautiful
----------------------------------------
Doc num: 1 Distance Score: 0.602
Doc: The sky is blue
```

BM25

```
Document Similarity Analysis using BM25
============================================================
Document 1 : The fox is definitely smarter than the dog
Top 2 similar docs:
----------------------------------------
Doc num: 8 BM25 Score: 7.334
Doc: The dog is smarter than the fox
----------------------------------------
Doc num: 7 BM25 Score: 3.88
Doc: The fox is quicker than the lazy dog
----------------------------------------

Document 2 : Java is a static typed programming language unlike Python
Top 2 similar docs:
----------------------------------------
Doc num: 5 BM25 Score: 5.501
Doc: Python and Java are popular Programming languages
----------------------------------------
Doc num: 6 BM25 Score: 4.586
Doc: Among Programming languages, both Python and Java are the most used in Analytics
----------------------------------------

Document 3 : I love to relax under the beautiful blue sky!
Top 2 similar docs:
----------------------------------------
Doc num: 2 BM25 Score: 7.334
Doc: The sky is blue and beautiful
```

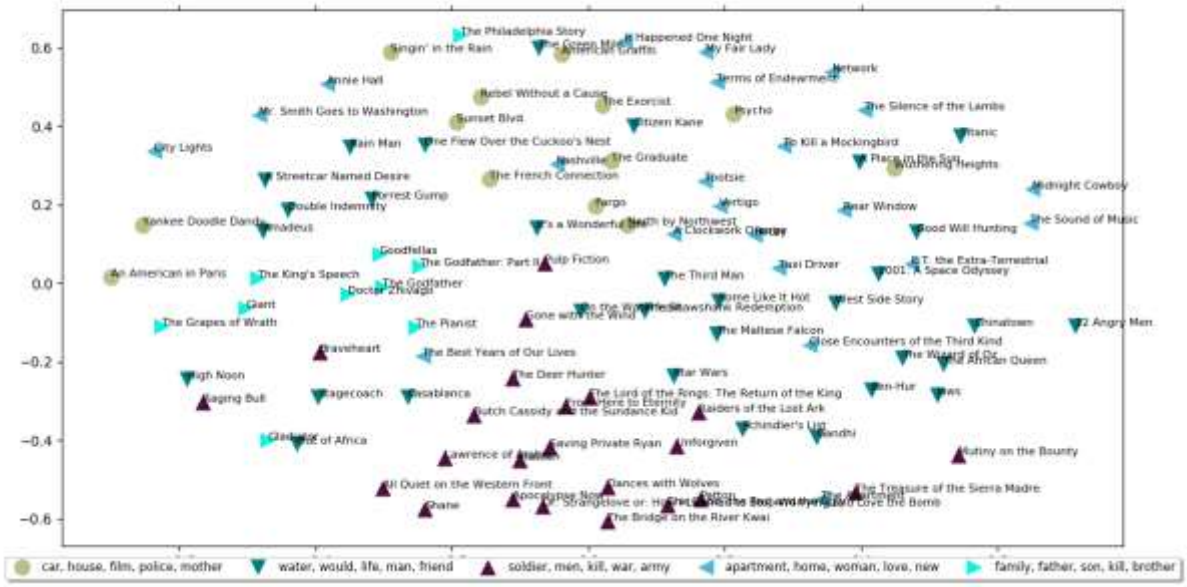
Document Clustering

K-means

Cluster details

```
Cluster 0 details:
--------------------
Key features: ['car', 'house', 'film', 'police', 'mother']
Movies in this cluster:
Psycho, Sunset Blvd., Singin' in the Rain, An American in Paris, The Exorcist, The French Connection, Fargo, The Graduate, American Graffiti, Wuthering Heights, Rebe
l Without a Cause, North by Northwest, Yankee Doodle Dandy
========================================
Cluster 1 details:
--------------------
Key features: ['water', 'would', 'life', 'man', 'friend']
Movies in this cluster:
The Shawshank Redemption, Schindler's List, Casablanca, One Flew Over the Cuckoo's Nest, Citizen Kane, The Wizard of Oz, Titanic, On the Waterfront, Forrest Gump, We
st Side Story, Star Wars, 2001: A Space Odyssey, Chinatown, It's a Wonderful Life, Some Like It Hot, 12 Angry Men, Amadeus, Gandhi, A Streetcar Named Desire, Ben-Hur
, Jaws, High Noon, A Place in the Sun, Rain Man, Out of Africa, Good Will Hunting, The Green Mile, The African Queen, Stagecoach, The Maltese Falcon, Double Indemnit
y, The Third Man
========================================
Cluster 2 details:
--------------------
Key features: ['soldier', 'men', 'kill', 'war', 'army']
Movies in this cluster:
Raging Bull, Gone with the Wind, Lawrence of Arabia, The Bridge on the River Kwai, Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb, Apocalypse N
ow, The Lord of the Rings: The Return of the King, From Here to Eternity, Saving Private Ryan, Unforgiven, Raiders of the Lost Ark, Patton, Braveheart, The Good, the
 Bad and the Ugly, Butch Cassidy and the Sundance Kid, The Treasure of the Sierra Madre, Platoon, Dances with Wolves, The Deer Hunter, All Quiet on the Western Front
, Shane, Pulp Fiction, Mutiny on the Bounty
========================================
```

Cluster visualization

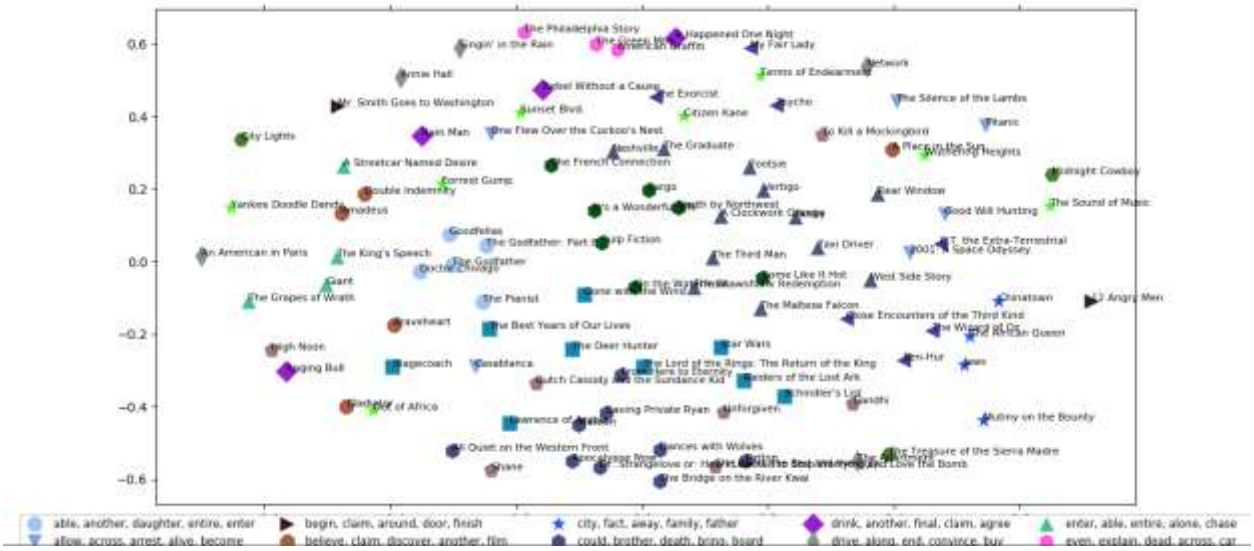


Affinity Propagation

Cluster details

```
Total Clusters: 17
Cluster 0 details:
--------------------
Key features: ['able', 'another', 'daughter', 'entire', 'enter']
Movies in this cluster:
The Godfather, The Godfather: Part II, Doctor Zhivago, The Pianist, Goodfellas
========================================
Cluster 1 details:
--------------------
Key features: ['allow', 'across', 'arrest', 'alive', 'become']
Movies in this cluster:
Casablanca, One Flew Over the Cuckoo's Nest, Titanic, 2001: A Space Odyssey, The Silence of the Lambs, Good Will Hunting
========================================
Cluster 2 details:
--------------------
Key features: ['appear', 'finish', 'fire', 'father', 'enough']
Movies in this cluster:
The Shawshank Redemption, Vertigo, West Side Story, Rocky, Tootsie, Nashville, The Graduate, The Maltese Falcon, A Clockwork Orange, Taxi Driver, Rear Window, The Th
ird Man
========================================
Cluster 3 details:
--------------------
Key features: ['arrive', 'finish', 'father', 'fire', 'everyone']
Movies in this cluster:
The Wizard of Oz, Psycho, E.T. the Extra-Terrestrial, My Fair Lady, Ben-Hur, The Exorcist, Close Encounters of the Third Kind
========================================
Cluster 4 details:
--------------------
Key features: ['begin', 'claim', 'around', 'door', 'finish']
Movies in this cluster:
12 Angry Men, Mr. Smith Goes to Washington
```

Cluster visualization



Hierarchical clustering

Dendrogram