CLUSTERING II

ALESSANDRO PANCONESI, SAPIENZA

From last time…

- More often than not, clustering problems are NP-hard and therefore (most likely) **computationally intractable**
- This is the case for instance for correlation clustering, k-means, k-median and k-center

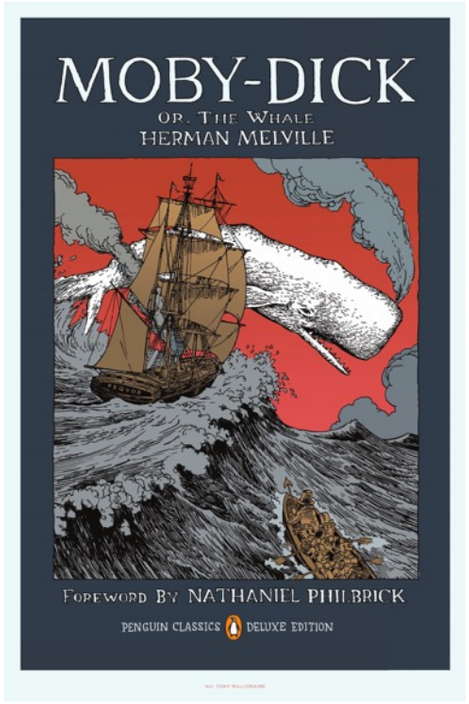
What to do?

- If it is important to find a “good” (optimal or near-optimal) solution one can use approximation algorithms with performance guarantee
- Oftentimes however, it is not so crucial to find near-optimal solutions and any “reasonable” one would do. In such cases, simple algorithms such as Lloyd, Ward or single-linkage, although very weak theoretically, work just fine

An example

Text Clustering

Bag of words

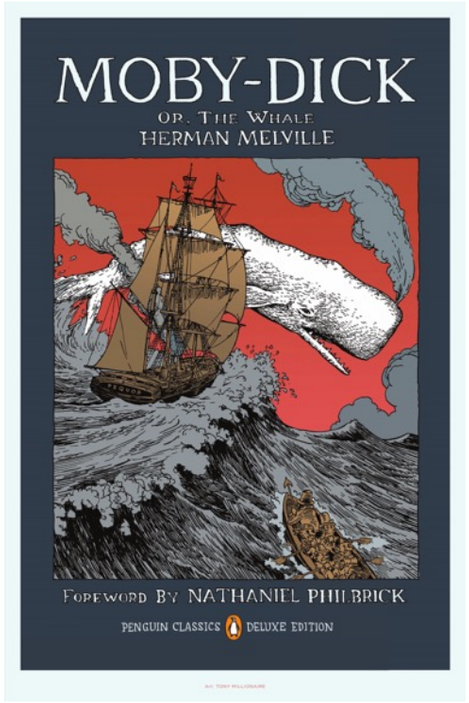


Call me Ishmael. Some years ago--never mind how long precisely—having little or no money in my purse, and nothing particular to interest me onshore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation…..

I = {call, me, Ishmael, some, …, the, circulation}

A text file (e.g a book) is identified with the set of words it contains.
This representation is crude but, as we will see, quite effective

Bag of words



Call me Ishmael. Some years ago--never mind how long precisely—having little or no money in my purse, and nothing particular to interest me onshore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation…..

about	ago	age	…	call	computer	…	internet	…	watery
1	1	0		1	0		0		1

Likewise, we can represent a document as a vector with as many coordinates as there are words in the language. Again, this is crude and more refined countings can be used, but this is already quite effective

Jaccard Similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard Similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

A = {I, had, pasta}

B = {the, pasta, looks, nice}

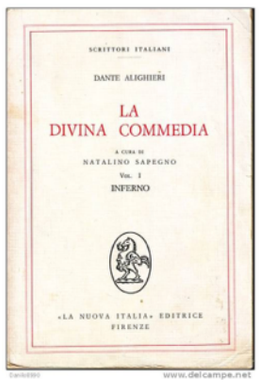
Jaccard Similarity

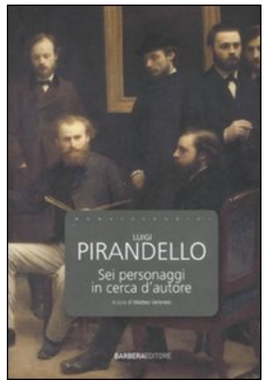
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{1}{6}$$

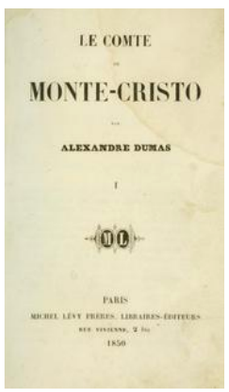
A = {I, had, pasta}

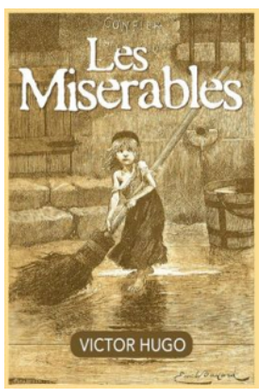
B = {the, pasta, looks, nice}

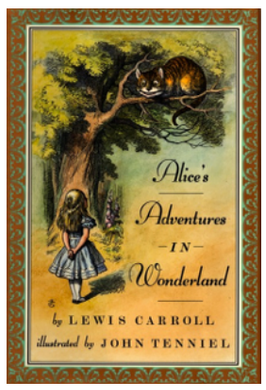
Our goal: Partitioning into languages

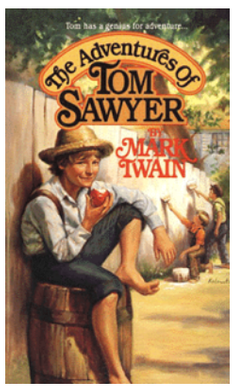


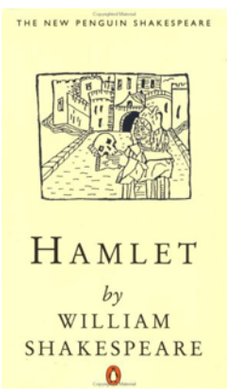




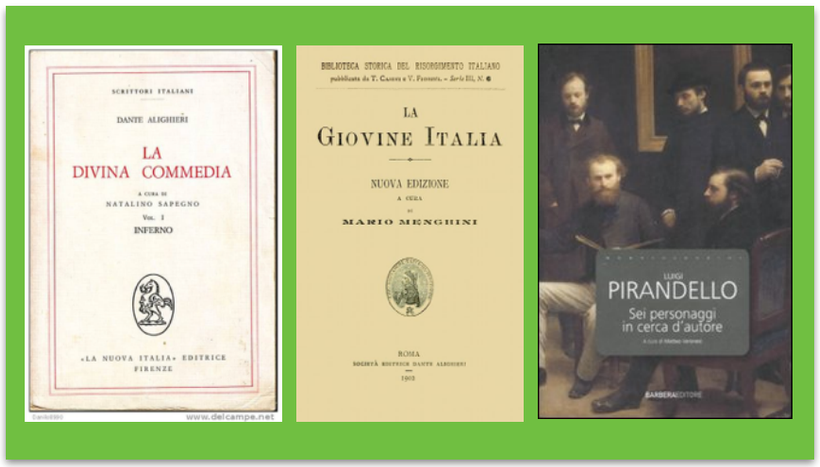


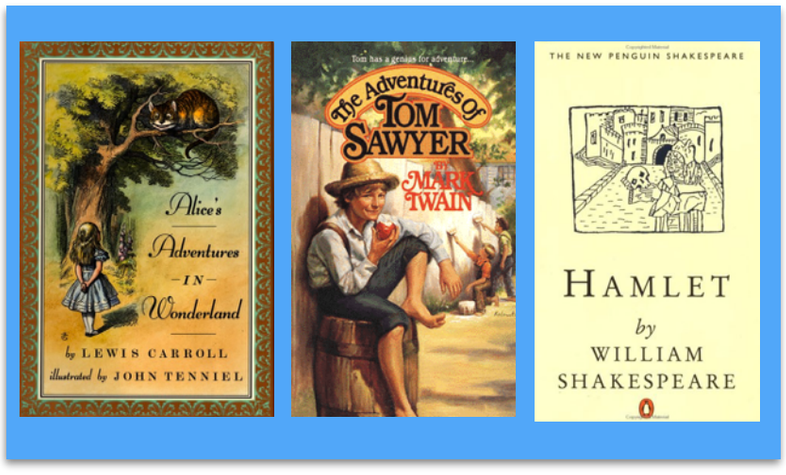






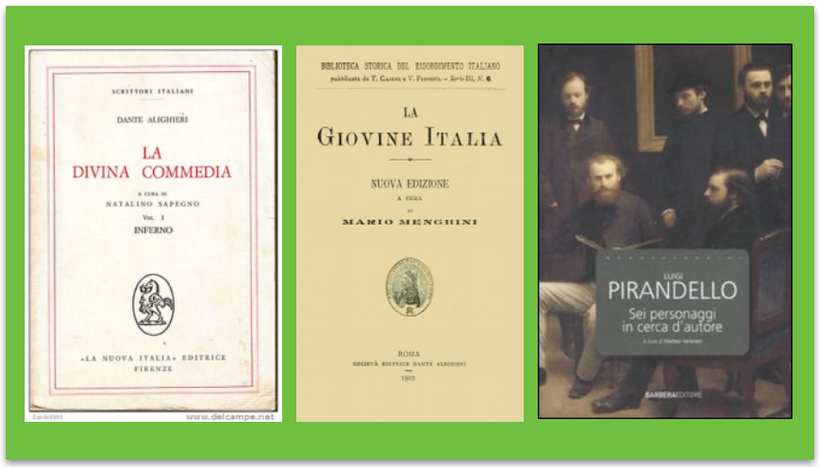
Partitioning into languages



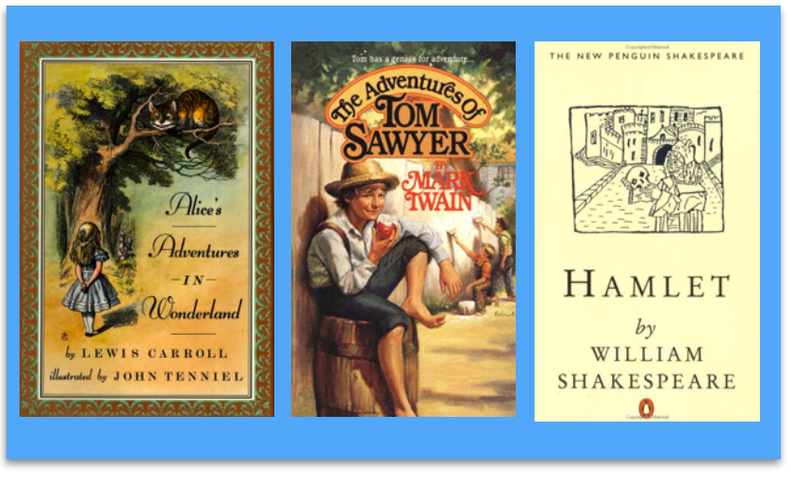


Partitioning into languages

Italian



French

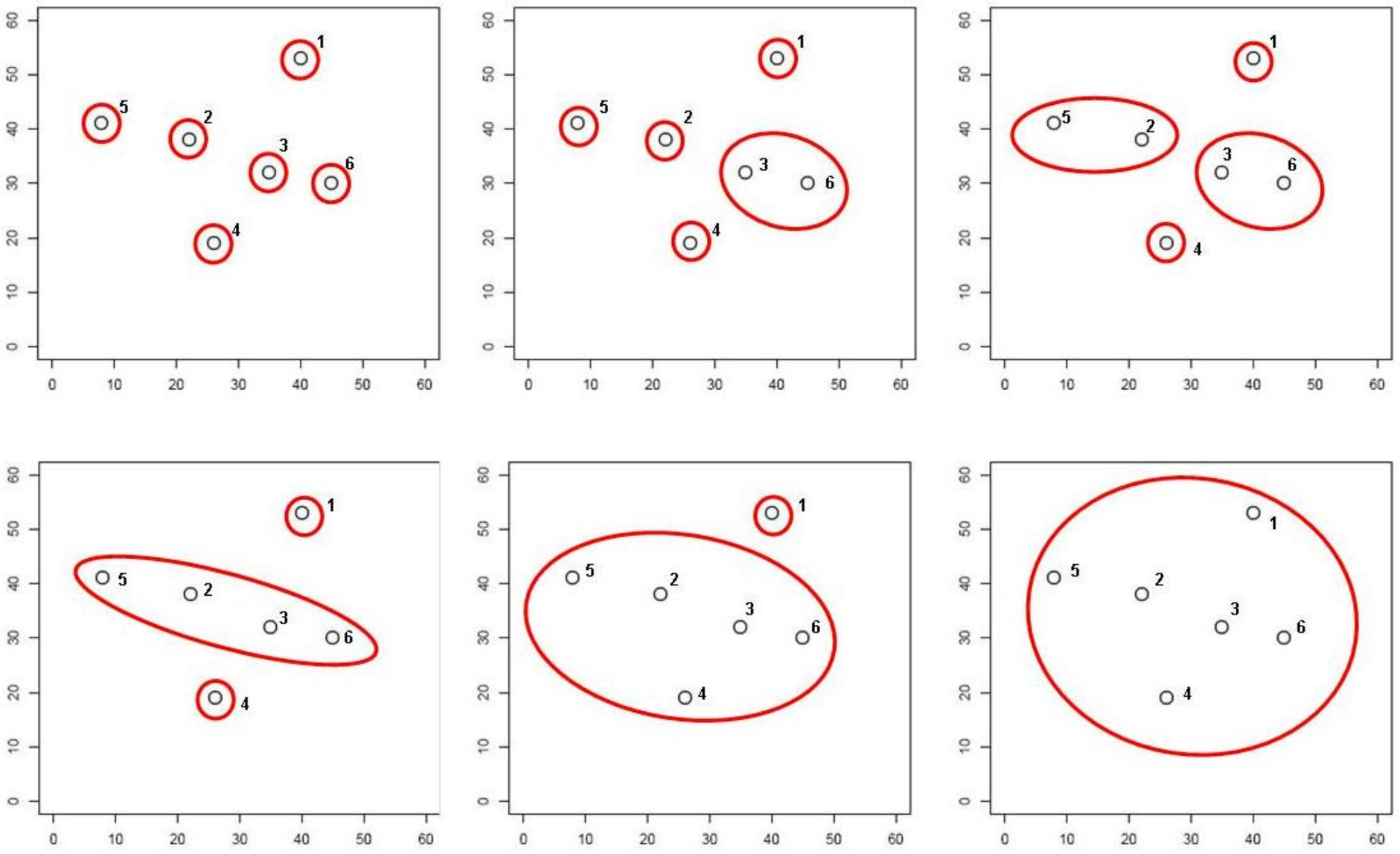


English

Single-Linkage algorithm

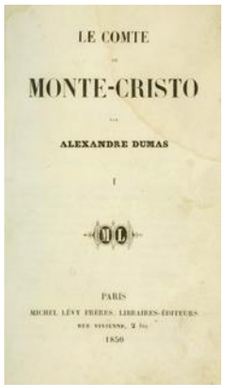
- Let P be the clustering into singletons of the data points.
- Until termination:
    - find the minimum distance between two points that are not in the same cluster;
    - join the two clusters containing those points.

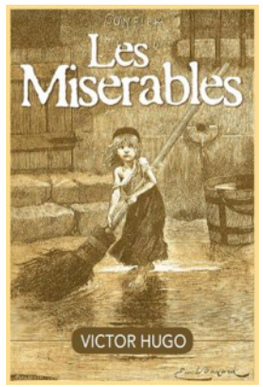
Single-Linkage algorithm

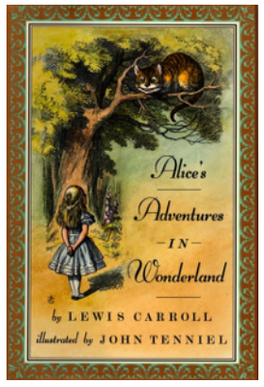


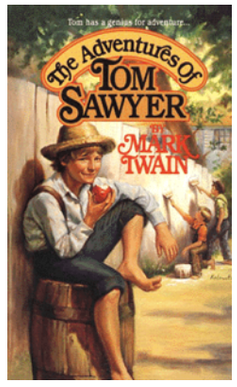
Single-Linkage algorithm

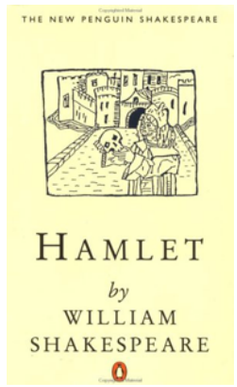
Jaccard Similarity

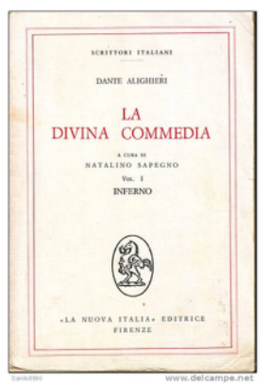








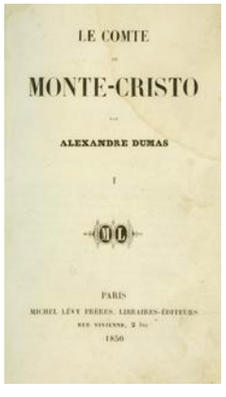


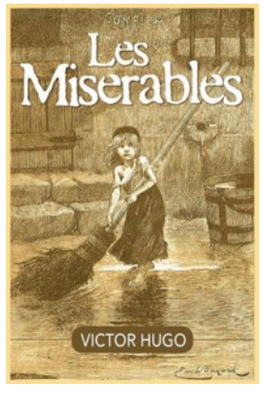


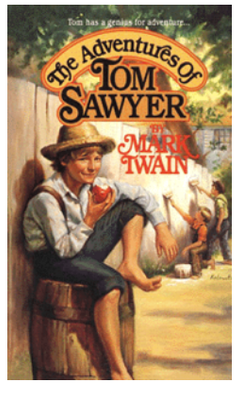
Single-Linkage algorithm

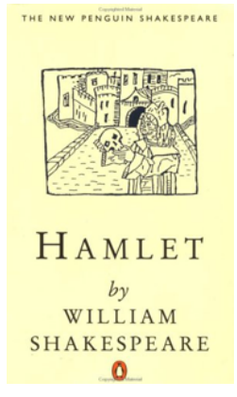
Jaccard Similarity

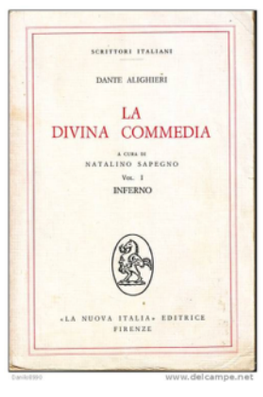
```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```









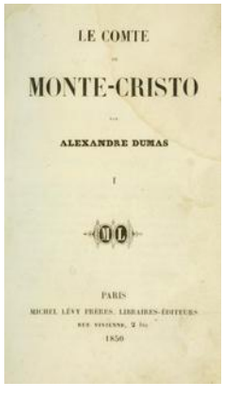


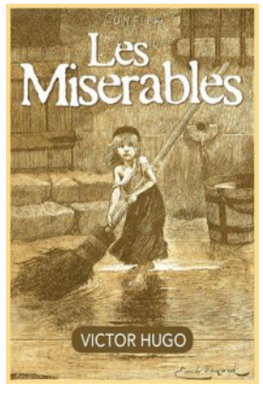
Single-Linkage algorithm

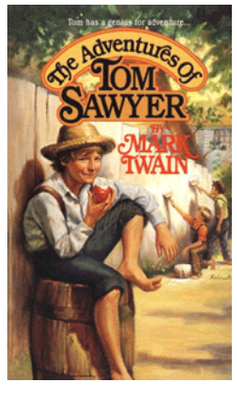
Jaccard Similarity

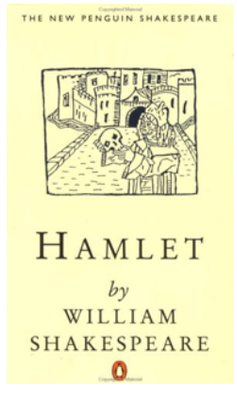
(1 - Jaccard) is a distance

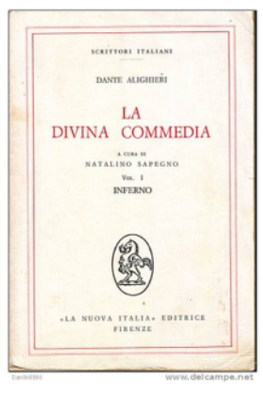
```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```

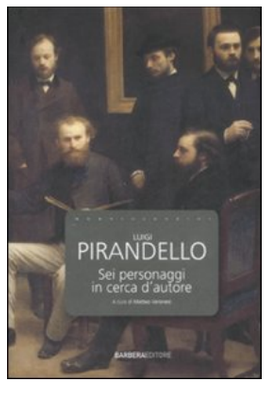






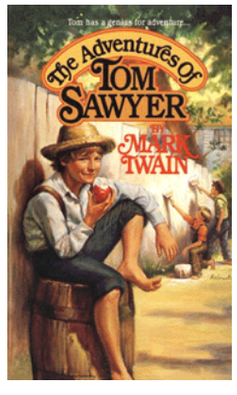


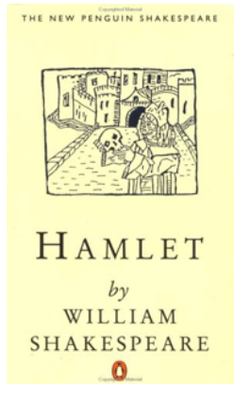


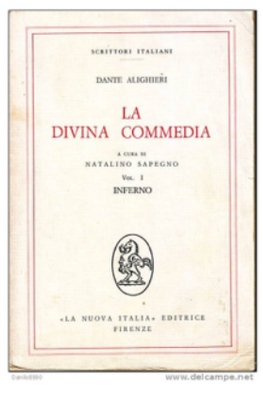


Single-Linkage algorithm

```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```

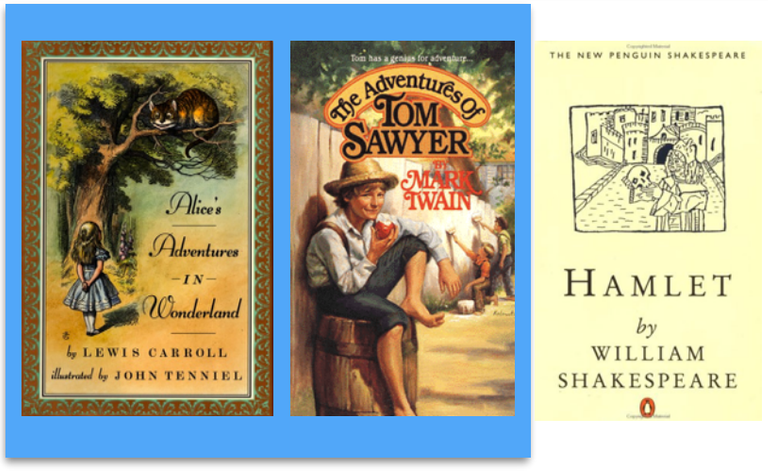






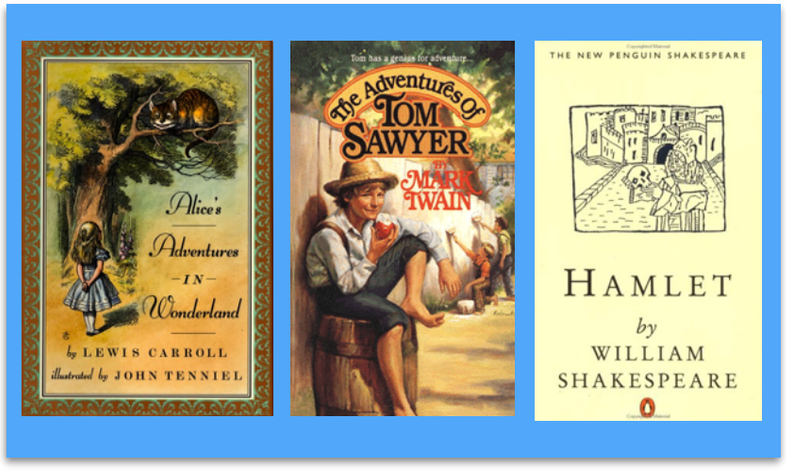
Single-Linkage algorithm

```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```



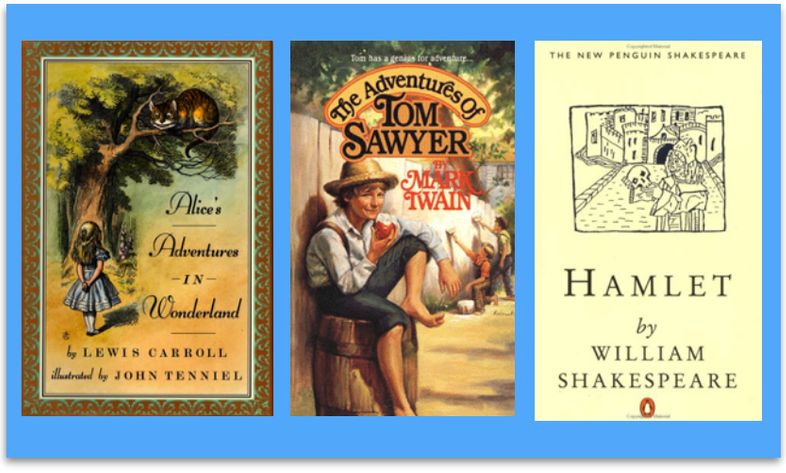
Single-Linkage algorithm

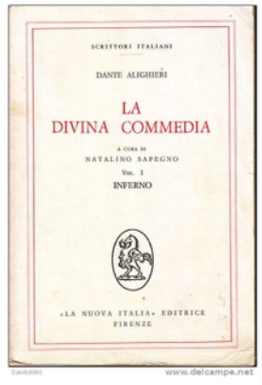
```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```

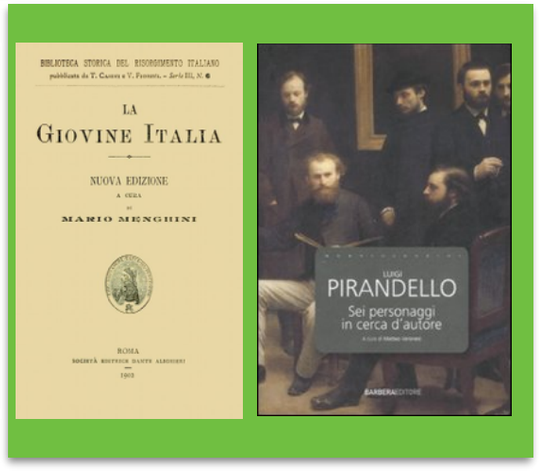


Single-Linkage algorithm

```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```

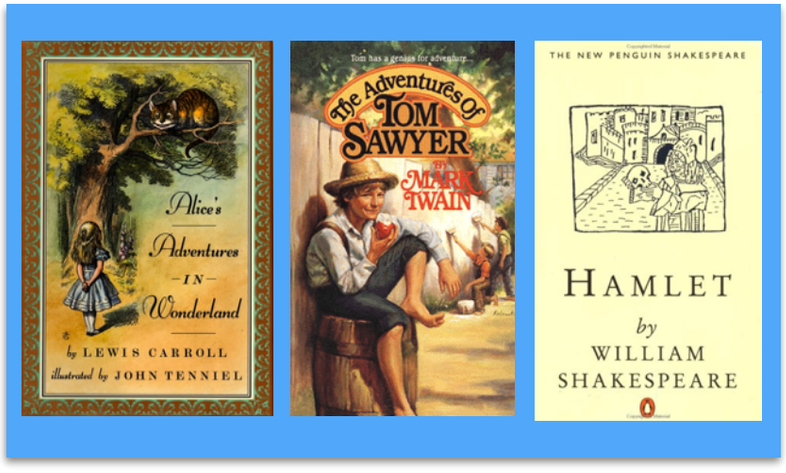






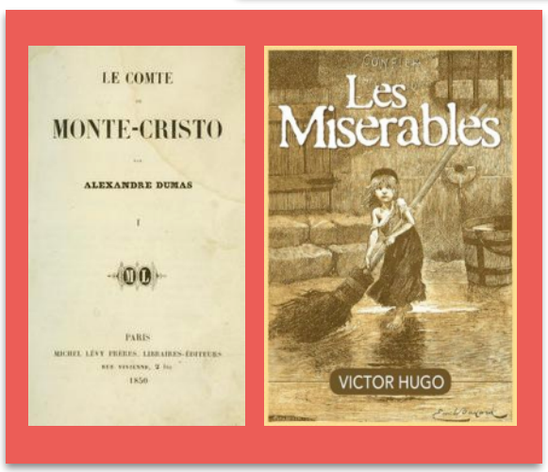
Single-Linkage algorithm

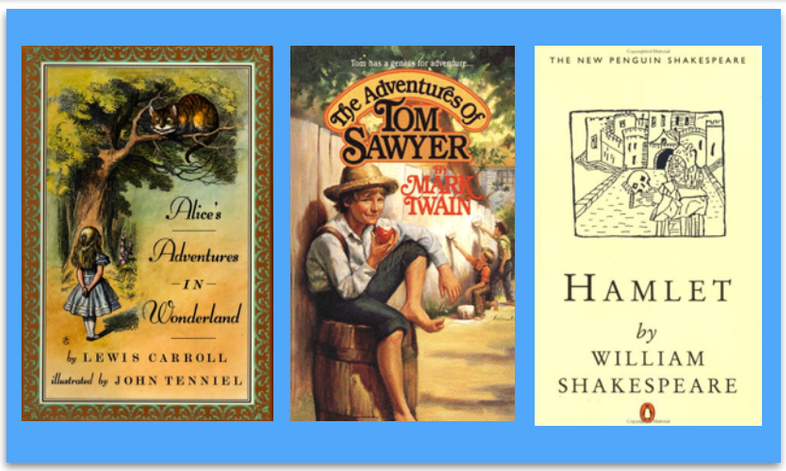
```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```

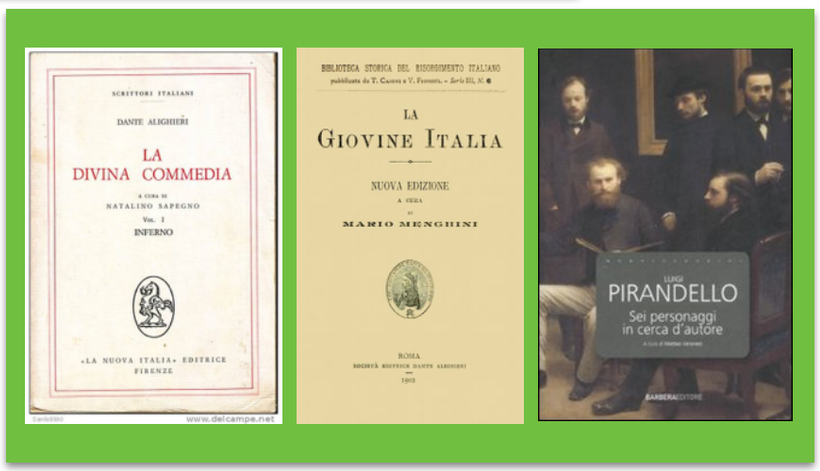


Single-Linkage algorithm

```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```



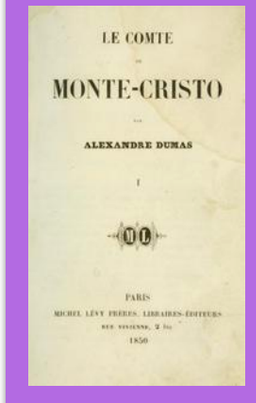


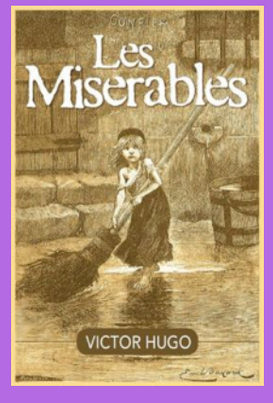


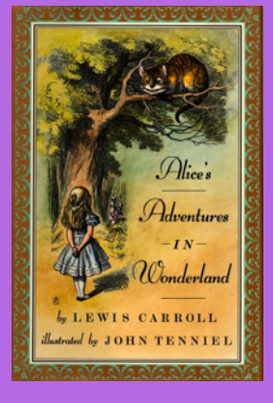
Perfect clustering into languages!

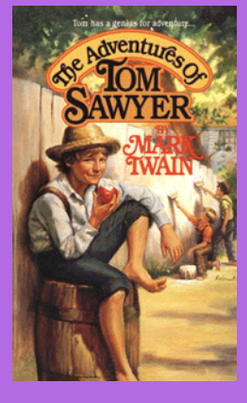
Single-Linkage algorithm

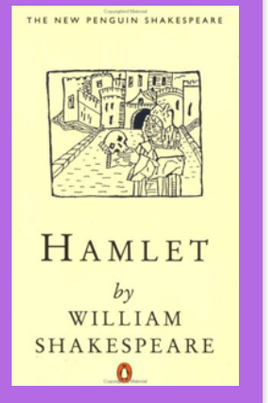
```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```

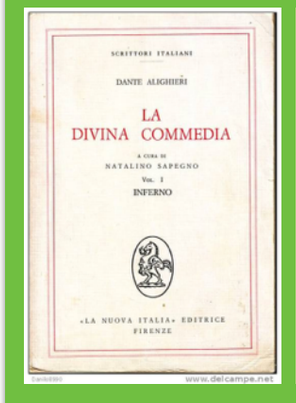






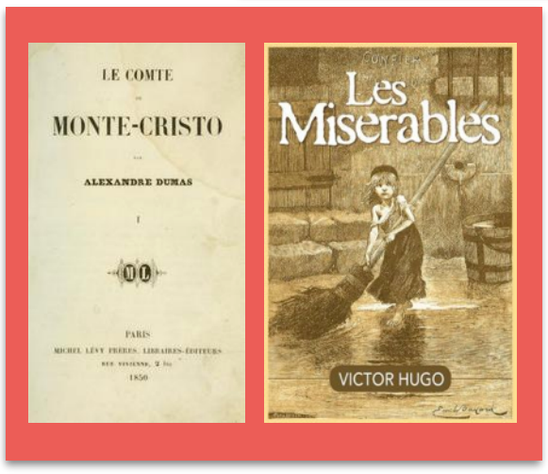


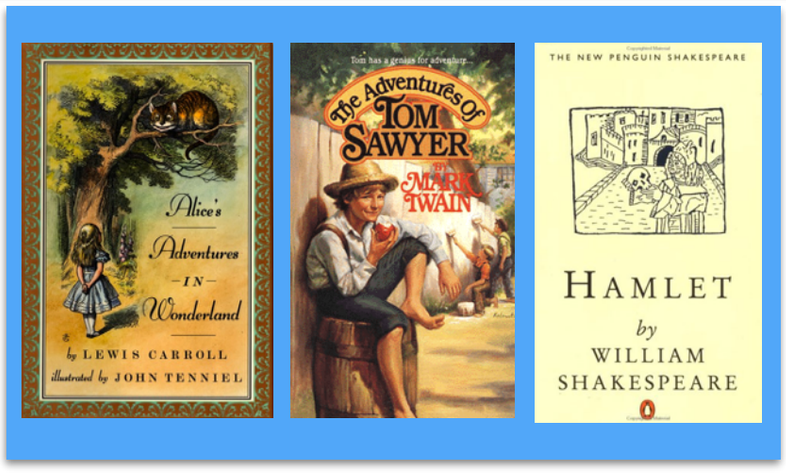


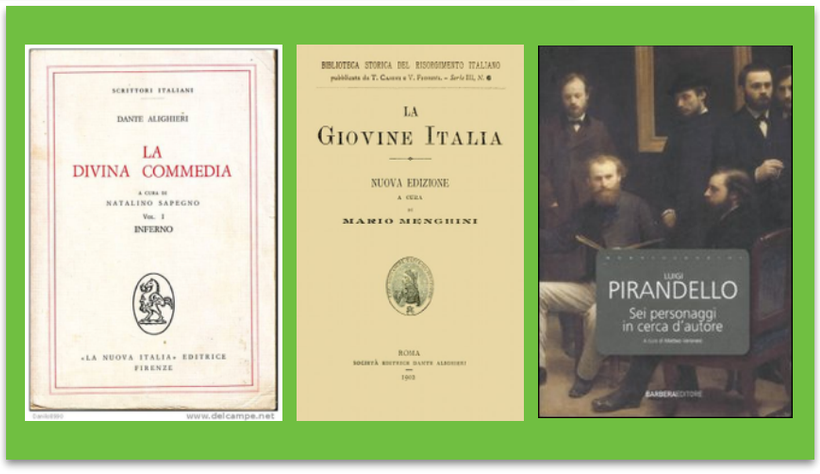


Single-Linkage algorithm

```
(flavio:{due}:lsh) ./lsh.py 1000 *txt
Jaccard( le_comte_de_monte_cristo.txt , les_miserables.txt ) = 0.229
Jaccard( alice_in_wonderland.txt , tom_sawyer.txt ) = 0.13
Jaccard( shakespeare_hamlet.txt , tom_sawyer.txt ) = 0.085
Jaccard( la_giovine_italia.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.083
Jaccard( alice_in_wonderland.txt , shakespeare_hamlet.txt ) = 0.082
Jaccard( dante_inferno.txt , la_giovine_italia.txt ) = 0.081
Jaccard( dante_inferno.txt , pirandello_sei_personaggi_in_cerca_dautore.txt ) = 0.072
Jaccard( le_comte_de_monte_cristo.txt , tom_sawyer.txt ) = 0.012
Jaccard( le_comte_de_monte_cristo.txt , shakespeare_hamlet.txt ) = 0.011
Jaccard( la_giovine_italia.txt , les_miserables.txt ) = 0.011
```







Perfect clustering into languages!

Single-Linkage algorithm

- Let P be the clustering into singletons of the data points.
- Until **termination**:
  - find the minimum distance between two points that are not in the same cluster;
  - join the two clusters containing those points.

Single-Linkage algorithm

- **Termination conditions**:
  - all the remaining distances are larger than *r,*
  - all the remaining distances are larger than the maximum distance times *0 < x < 1,*
  - there are *k* clusters left.

Single-Linkage algorithm

- **Termination conditions**: *[Kleinberg, '02]*
  - all the remaining distances are larger than *r*,
    Richness and Consistency
  - all the remaining distances are larger than the maximum distance times *0 < x < 1*,
    Richness and Scale Invariance
  - there are *k* clusters left.
    Consistency and Scale Invariance

Single-Linkage is too slow for large datasets

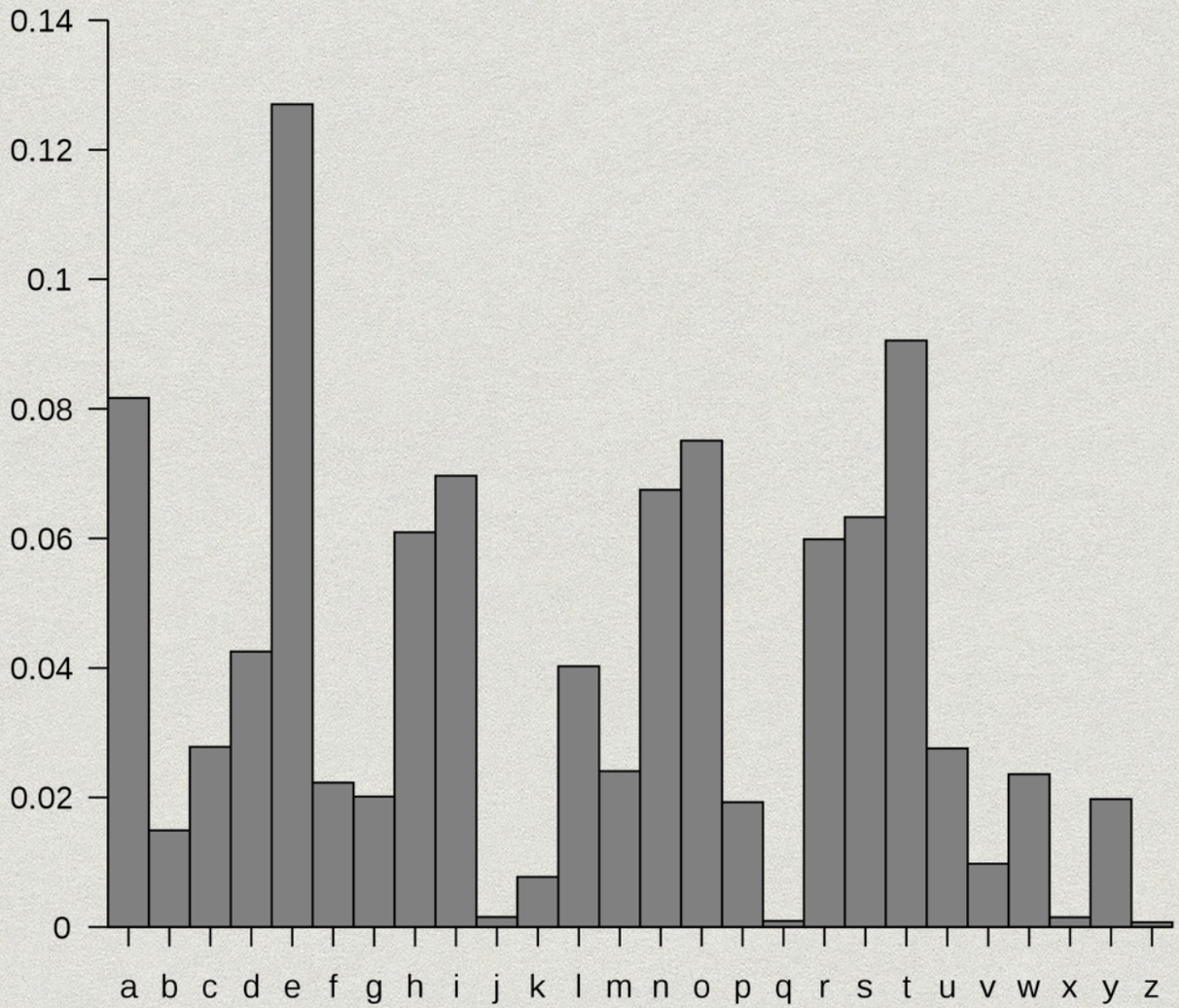
- Even just computing all the n^2 distances is prohibitive: think of 10^6, or 10^9, data points.

- One needs smarter (almost-linear) ways to deal with such n's.

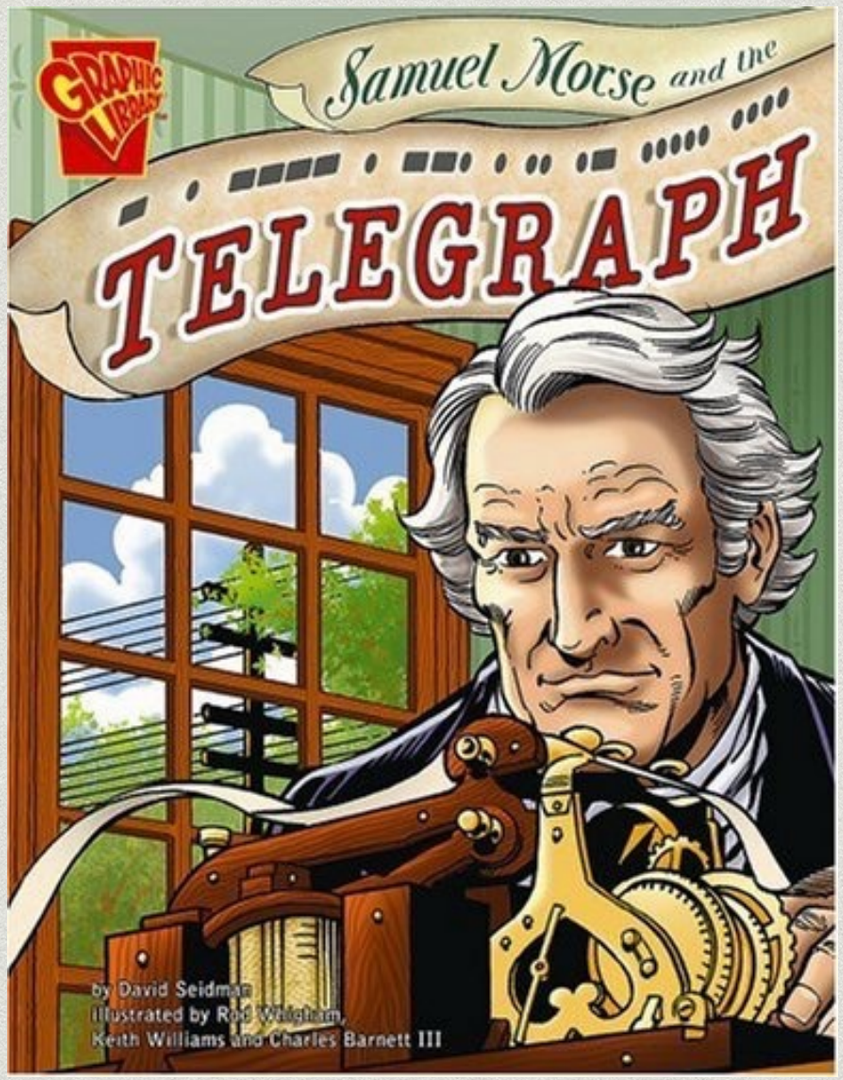
Single-linkage, take 2

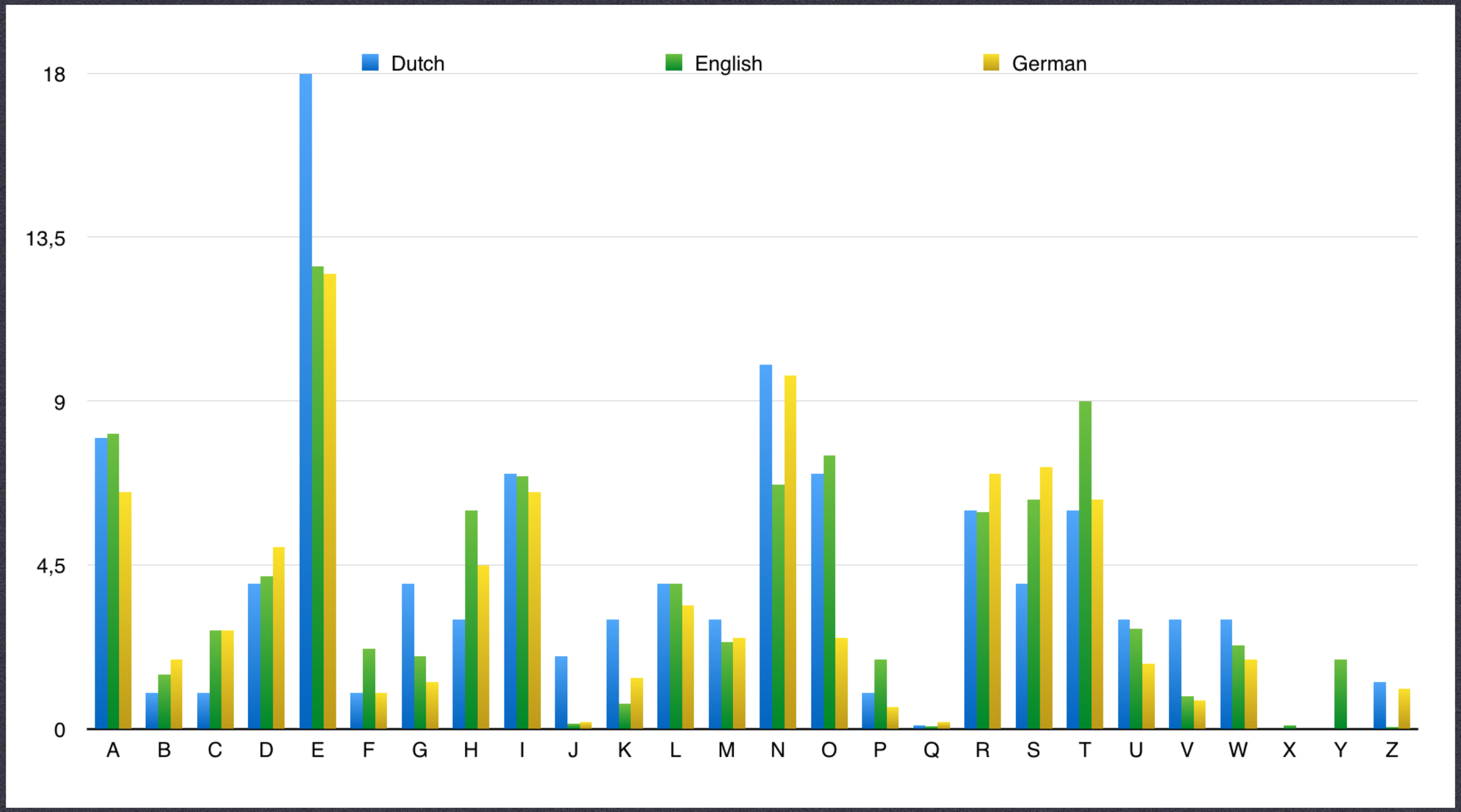
- We will try now a different approach, using the vector representation…

Letter frequency



English

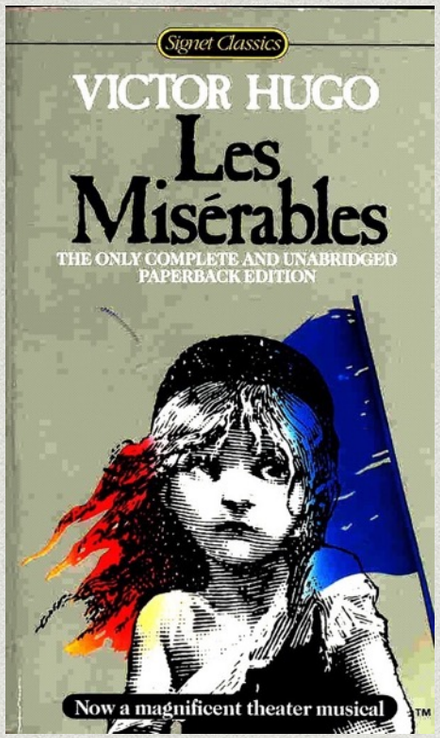




LETTER FREQUENCIES

DUTCH VS ENGLISH VS GERMAN

Single-linkage, take 2

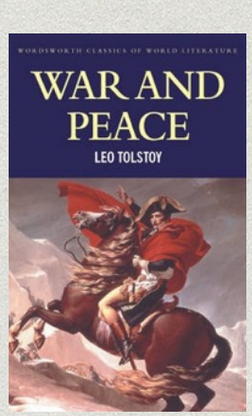


[0.085, 0.01, 0.031, 0.036, 0.153, 0.012, 0.009, 0.01, 0.079, 0.006, 0.0, 0.061, 0.031, 0.07, 0.053, 0.027, 0.013, 0.064, 0.076, 0.078, 0.065, 0.021, 0.0, 0.004, 0.004, 0.002]

a b c d e f g h i j k l m n o p q r s t u v w x y z

Single-linkage, take 2

[0.085, 0.01, 0.031, 0.036, 0.153, 0.012, …, 0.0, 0.004, 0.004, 0.002]

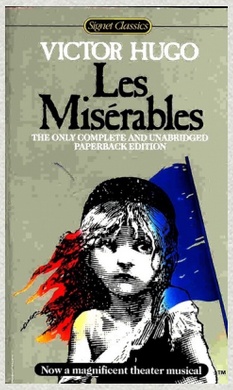


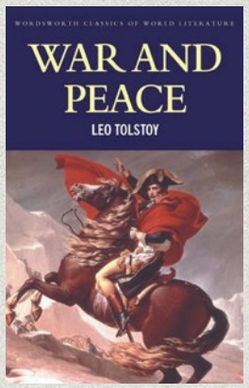
[0.081, 0.014, 0.024, 0.047, 0.125, …, 0.011, 0.023, 0.002, 0.018, 0.001]

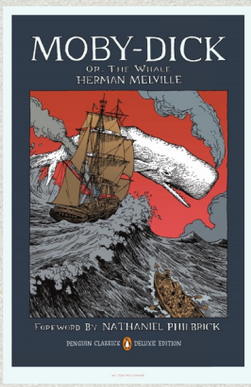
[0.108, 0.008, 0.054, 0.036, 0.117, 0.013, …, 0.02, 0.0, 0.0, 0.0, 0.004]

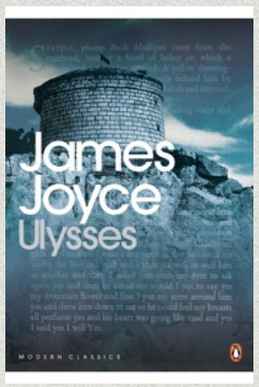
Single-linkage, take 2

```
0.0136 MobyDick-Ulysses
0.0137 MobyDick-WarAndPeace
0.0166 Ulysses-WarAndPeace
0.0169 LesMiserables-MonteCristo
0.02 Inferno-SeiPersonaggi
0.0797 Inferno-MonteCristo
0.0832 Inferno-LesMiserables
0.0883 Inferno-Ulysses
0.0884 MonteCristo-Ulysses
0.0891 LesMiserables-Ulysses
0.0904 LesMiserables-MobyDick
0.0912 MobyDick-MonteCristo
```



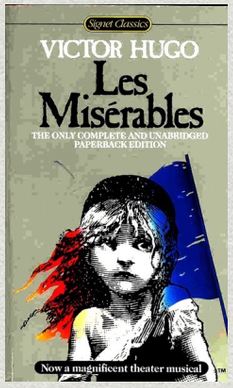


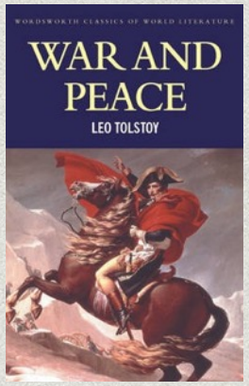


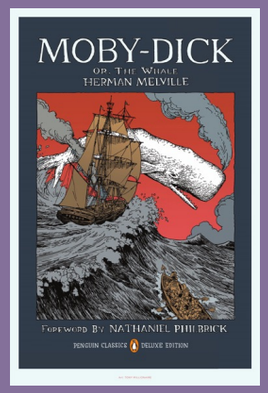


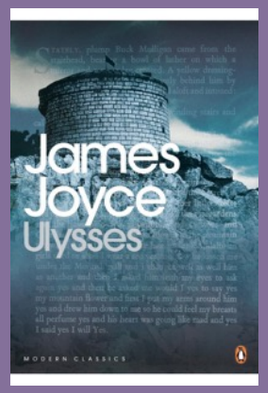
Single-linkage, take 2

```
0.0136 MobyDick-Ulysses
0.0137 MobyDick-WarAndPeace
0.0166 Ulysses-WarAndPeace
0.0169 LesMiserables-MonteCristo
0.02 Inferno-SeiPersonaggi
0.0797 Inferno-MonteCristo
0.0832 Inferno-LesMiserables
0.0883 Inferno-Ulysses
0.0884 MonteCristo-Ulysses
0.0891 LesMiserables-Ulysses
0.0904 LesMiserables-MobyDick
0.0912 MobyDick-MonteCristo
```



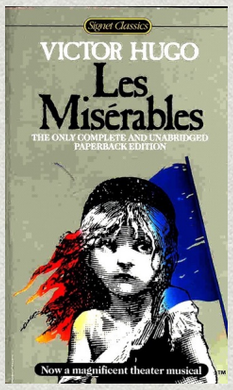


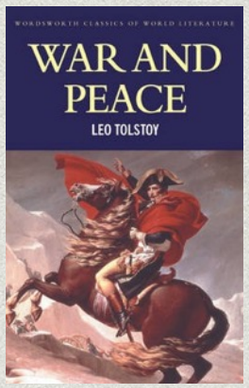


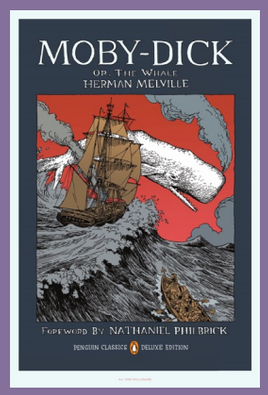


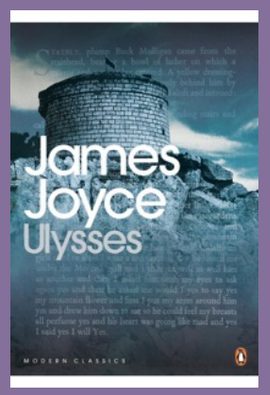
Single-linkage, take 2

```
0.0136 MobyDick-Ulysses
0.0137 MobyDick-WarAndPeace
0.0166 Ulysses-WarAndPeace
0.0169 LesMiserables-MonteCristo
0.02 Inferno-SeiPersonaggi
0.0797 Inferno-MonteCristo
0.0832 Inferno-LesMiserables
0.0883 Inferno-Ulysses
0.0884 MonteCristo-Ulysses
0.0891 LesMiserables-Ulysses
0.0904 LesMiserables-MobyDick
0.0912 MobyDick-MonteCristo
```



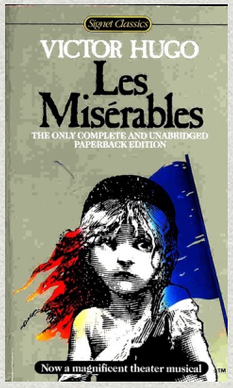


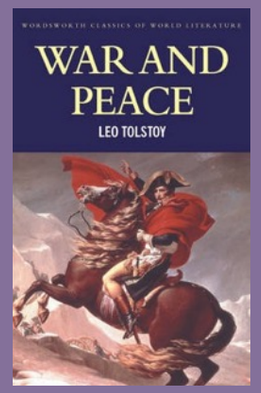


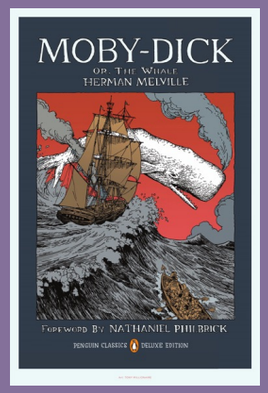


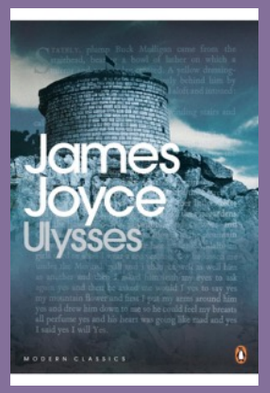
Single-linkage, take 2

```
0.0136 MobyDick-Ulysses
0.0137 MobyDick-WarAndPeace
0.0166 Ulysses-WarAndPeace
0.0169 LesMiserables-MonteCristo
0.02 Inferno-SeiPersonaggi
0.0797 Inferno-MonteCristo
0.0832 Inferno-LesMiserables
0.0883 Inferno-Ulysses
0.0884 MonteCristo-Ulysses
0.0891 LesMiserables-Ulysses
0.0904 LesMiserables-MobyDick
0.0912 MobyDick-MonteCristo
```



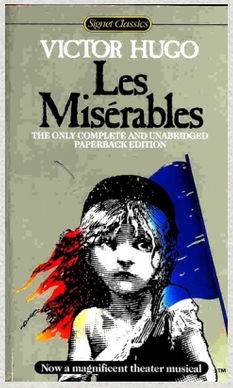


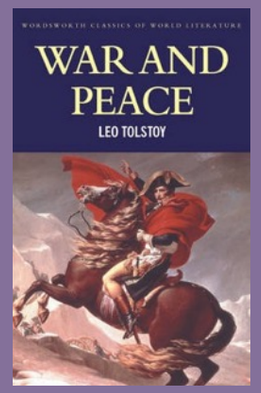


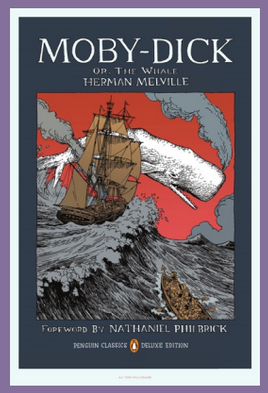


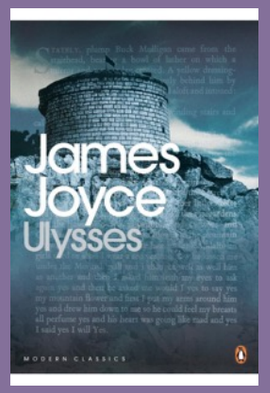
Single-linkage, take 2

```
0.0136 MobyDick-Ulysses
0.0137 MobyDick-WarAndPeace
0.0166 Ulysses-WarAndPeace
0.0169 LesMiserables-MonteCristo
0.02 Inferno-SeiPersonaggi
0.0797 Inferno-MonteCristo
0.0832 Inferno-LesMiserables
0.0883 Inferno-Ulysses
0.0884 MonteCristo-Ulysses
0.0891 LesMiserables-Ulysses
0.0904 LesMiserables-MobyDick
0.0912 MobyDick-MonteCristo
```



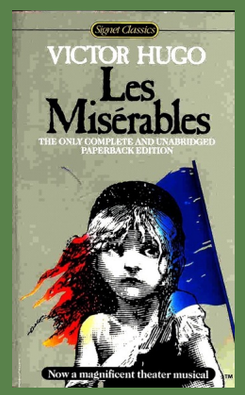


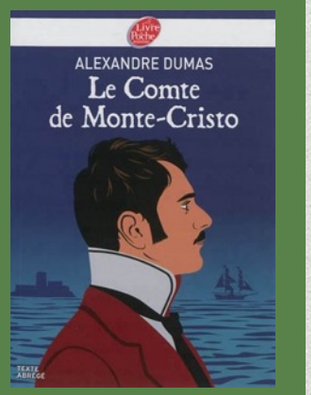


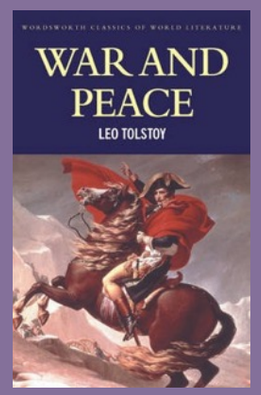


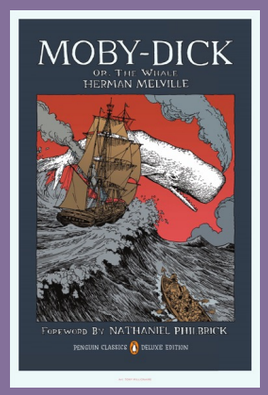
Single-linkage, take 2

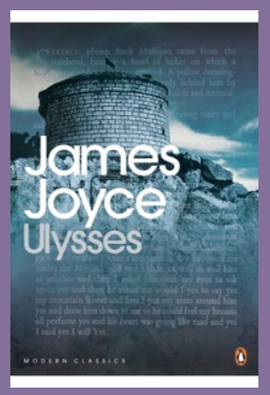
```
0.0136 MobyDick-Ulysses
0.0137 MobyDick-WarAndPeace
0.0166 Ulysses-WarAndPeace
0.0169 LesMiserables-MonteCristo
0.02 Inferno-SeiPersonaggi
0.0797 Inferno-MonteCristo
0.0832 Inferno-LesMiserables
0.0883 Inferno-Ulysses
0.0884 MonteCristo-Ulysses
0.0891 LesMiserables-Ulysses
0.0904 LesMiserables-MobyDick
0.0912 MobyDick-MonteCristo
```







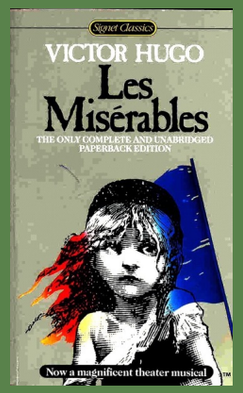


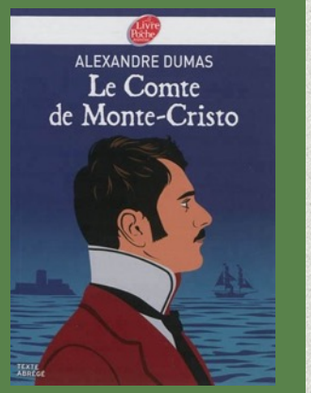


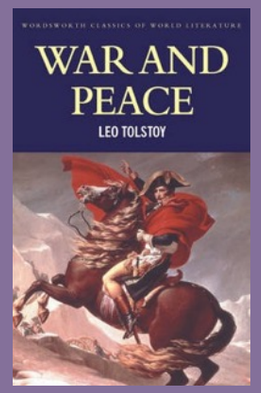
Single-linkage, take 2

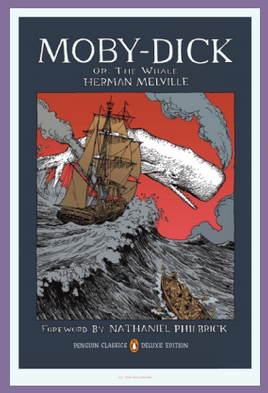
```
0.0136 MobyDick-Ulysses
0.0137 MobyDick-WarAndPeace
0.0166 Ulysses-WarAndPeace
0.0169 LesMiserables-MonteCristo
0.02 Inferno-SeiPersonaggi
0.0797 Inferno-MonteCristo
0.0832 Inferno-LesMiserables
0.0883 Inferno-Ulysses
0.0884 MonteCristo-Ulysses
0.0891 LesMiserables-Ulysses
0.0904 LesMiserables-MobyDick
0.0912 MobyDick-MonteCristo
```

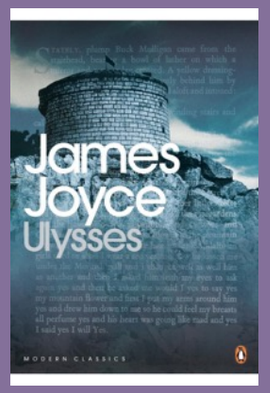
Perfect clustering

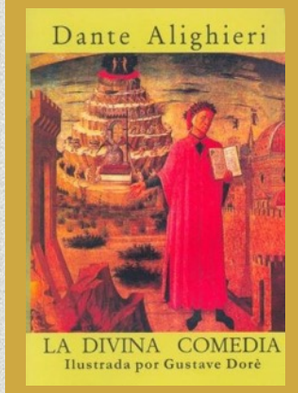






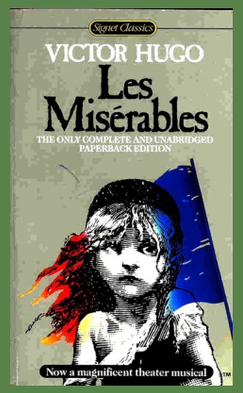


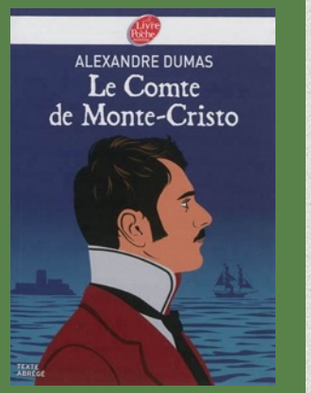


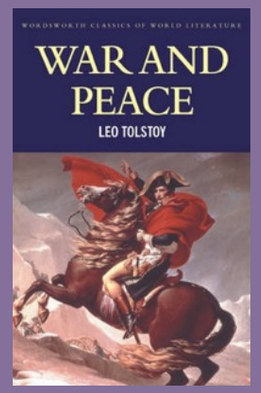


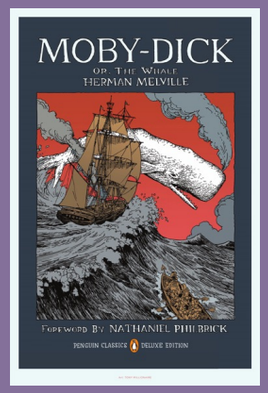
Single-linkage, take 2

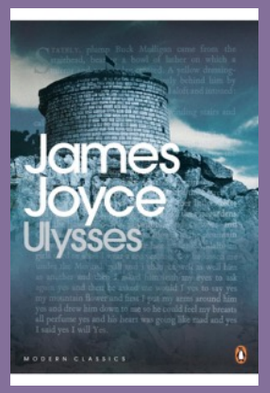
```
0.0136 MobyDick-Ulysses
0.0137 MobyDick-WarAndPeace
0.0166 Ulysses-WarAndPeace
0.0169 LesMiserables-MonteCristo
0.02 Inferno-SeiPersonaggi
0.0797 Inferno-MonteCristo
0.0832 Inferno-LesMiserables
0.0883 Inferno-Ulysses
0.0884 MonteCristo-Ulysses
0.0891 LesMiserables-Ulysses
0.0904 LesMiserables-MobyDick
0.0912 MobyDick-MonteCristo
```

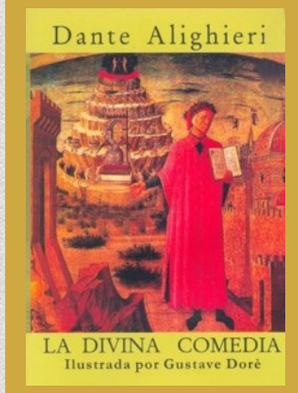


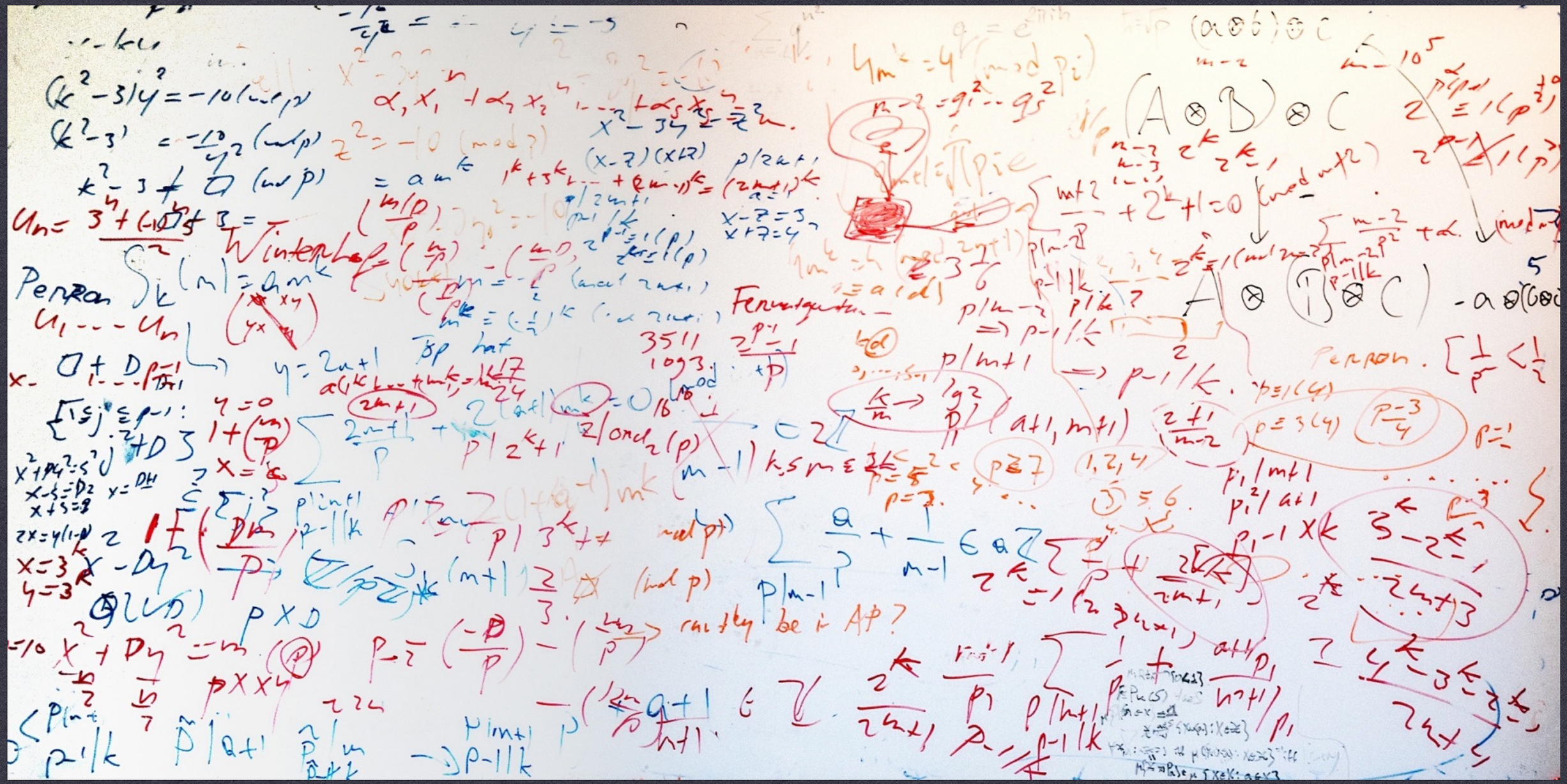












BACK TO THE BOARD