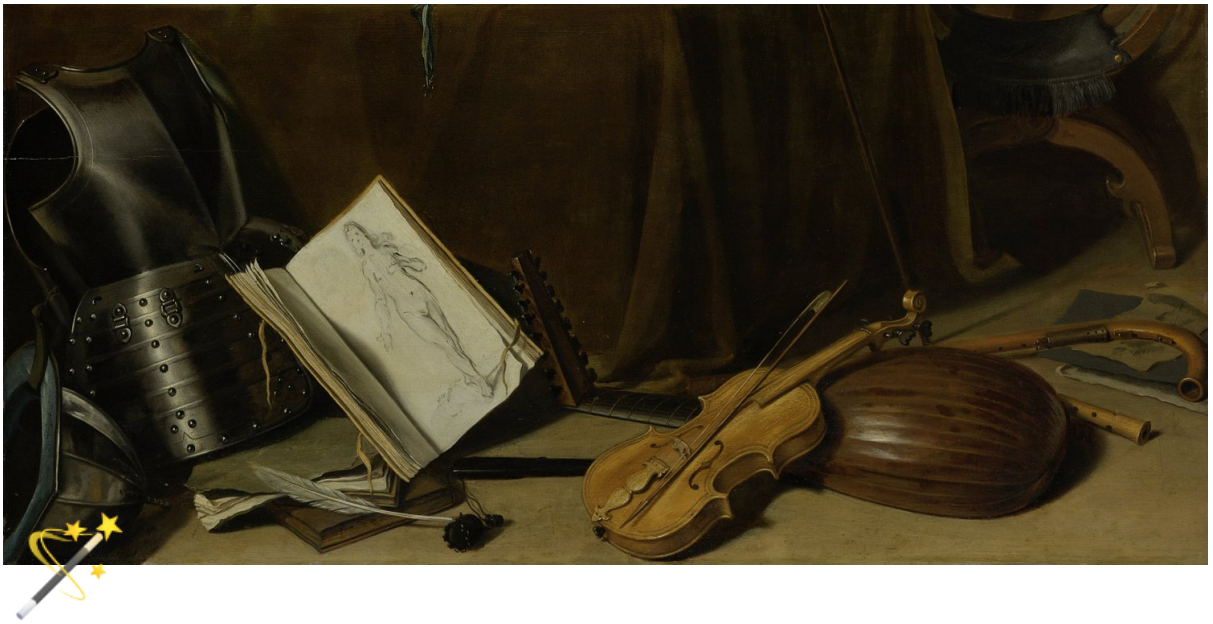

HW1 - Finding Similar Items

Solution

The solution I have built uses a few classes, which are described below.

Main

Ties all the other classes together and parses their outputs into each other when needed. Has some basic methods that call on the other classes, like readAllDocs() and createShingles().

Shingling

The Shingling class is an instance of all the shingles for a document. It takes the document's content and transforms it into shingles of length k (set to 10 in the current implementation). This is stored in a set as plain text.

It also creates a set of hashed shingles, which is later used in the implementation in order to save memory.

Each instance of the singling class is stored in its respective instance of the doc class.

MinHash

The MinHash class takes an input of all the documents and creates a characteristic matrix where each column represents a document. The values of each element in

the column correspond to if a specific shingle is present in that document.

This is used to create a signature matrix, which is a smaller matrix where each column is a kind of signature for the document. If two columns are similar, their respective document is probably also similar.

LSH

The LSH (Locality-Sensitive Hashing) class is implemented to be able to “skim through” all the documents in order to find similar ones. Instead of having to compare every document to each other, we split them into bands and hash each band of each column (i.e. each document) and put them into so-called buckets corresponding to the generated hash value. Since it is unlikely for two different elements in a band to give the same hash value, we assume that the two elements in the band are the same if they get hashed into the same bucket.

Now, we can go through all the buckets and see the sets of documents that the algorithm found to be similar in at least one of the bands and compare them in the traditional Jaccard-similarity way to check for false positives.

After this check, we output the documents which are similar according to our algorithm.

(Be mindful that we might miss some similar documents using LSH. We trade this for better time complexity).

Other classes

- **doc:** stores the content and the shingles of a document
- **CompareSets**: Used to compare the Jaccard-similarity of a set of shingles. Can be found in shingling.py

How to build and run

The code is run by simply running the main.py file. Python 3.10 was used when testing, but other versions might work.

Run this to get all the required libraries: pip install -r requirements.txt.

Change the self.dir variable to change the data set used. All the files in the given folder will be read and compared to each other.

It then prints all the pairs of documents that were found to be similar, as well as their Jaccard similarity.

Results

> 💡 For results, your report should include results from running with the specific inputs of the assignment (if specified in the instructions; otherwise, results from running sample inputs should be included). The report should also include info on how long it took your implementation to compute each set of the results.

Test 1

```
self.k_shingle = 10     # Shingle size
self.k_perm = 100       # Permutations
self.t = 0.8            # Threshold
self.b = 5              # Bands
```

Biden.txt and BidenWithNoise.txt are similar with Jaccard-similarity 0.94
Obama.txt and ObamaChangedToClinton.txt are similar with Jaccard-similarity 0.91
2 similar docs found

---------- FACIT ----------
('Biden.txt', 'BidenWithNoise.txt') 0.9374570446735395
('Obama.txt', 'ObamaChangedToClinton.txt') 0.8752551020408164

Time: 0.2194666862487793

> 💡 Here we find the two datasets that are similar to each other.
> The facit shows the true values, where we can see that our estimations are pretty accurate

Test 2: Lower threshold

```
self.k_shingle = 10     # Shingle size
self.k_perm = 100       # Permutations
```

```
self.t = 0.5            # Threshold
self.b = 5              # Bands
```

Obama.txt and ObamaChangedToClinton.txt are similar with Jaccard-similarity 0.87
Biden.txt and BidenWithNoise.txt are similar with Jaccard-similarity 0.92
2 similar docs found

---------- FACIT ----------

('Biden.txt', 'BidenTrump.txt') 0.5191116805466581
('Biden.txt', 'BidenWithNoise.txt') 0.9374570446735395
('BidenTrump.txt', 'BidenWithNoise.txt') 0.5382340515420363
('Obama.txt', 'ObamaChangedToClinton.txt') 0.8752551020408164

Time: 0.22698473930358887

> 💡 If we lower the threshold, we can see that we still miss some documents that are below the threshold, but are only 50% similar. The reason for this is that they are different enough that the none of the 5 bands are not exactly equal to each other

Test 3: Lower threshold and higher amounts of bands

```
self.k_shingle = 10     # Shingle size
self.k_perm = 100       # Permutations
self.t = 0.5            # Threshold
self.b = 50             # Bands
```

Biden.txt and BidenWithNoise.txt are similar with Jaccard-similarity 0.95
Obama.txt and ObamaChangedToClinton.txt are similar with Jaccard-similarity 0.88
Biden.txt and BidenTrump.txt are similar with Jaccard-similarity 0.52
BidenTrump.txt and BidenWithNoise.txt are similar with Jaccard-similarity 0.52
4 similar docs found

---------- FACIT ----------
('Biden.txt', 'BidenTrump.txt') 0.5191116805466581
('Biden.txt', 'BidenWithNoise.txt') 0.9374570446735395

('BidenTrump.txt', 'BidenWithNoise.txt') 0.5382340515420363
('Obama.txt', 'ObamaChangedToClinton.txt') 0.8752551020408164

Time: 0.22805500030517578

> 💡 If we make more bands, meaning we only have 2 rows in each we find these. Some experiments using different band sizes could lead to finding a smaller, but still effective hyperparameter

Test 4: Spam dataset

SMS1124.txt and SMS1133.txt are similar with Jaccard-similarity 1.0
SMS1132.txt and SMS1152.txt are similar with Jaccard-similarity 1.0
SMS1134.txt and SMS116.txt are similar with Jaccard-similarity 1.0
3 similar docs found

---------- FACIT ----------
('SMS1124.txt', 'SMS1133.txt') 1.0
('SMS1132.txt', 'SMS1152.txt') 1.0
('SMS1134.txt', 'SMS116.txt') 1.0
Time: 4.434859752655029

> 💡 This finds duplicate SMSes, which is an indication of spam

Test 5: Movies

```
self.k_shingle = 10     # Shingle size
self.k_perm = 100       # Permutations
self.t = 0.2            # Threshold
self.b = 50             # Bands
```

Similarity for SW_4.txt and SW_5.txt **below threshold**: 0.12
Similarity for SW_4.txt and SW_5.txt **below threshold**: 0.12
LSH created
No similar docs found

---------- FACIT ----------

No similar documents found

Time: 11.696559190750122

> 💡 Not really interesting since none of the movie scripts are similar. However, it showcases that the time complexity is managable

Datasets

Presidents Dataset

DataPresidents is each American president's opening paragraphs on Wikipedia

All presidents from Biden to Bush Senior were used, as well as some modified documents, see below.

This is the main dataset that is according to the instructions, but I added some others for the fun of it.


Joe Biden

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

"Joseph Biden" and "Biden" redirect here. For his son, Joseph Biden III, see Beau Biden. For other uses, see Biden (disambiguation).

Joseph Robinette Biden Jr. (/ˈbaɪdən/ BY-dən; born November 20, 1942) is an American politician who is the 46th and current president of the United States. Ideologically a moderate member of the Democratic Party, he previously served as the 47th vice president from 2009 to 2017 under President Barack Obama and represented Delaware in the United States Senate from 1973 to 2009.

Born in Scranton, Pennsylvania, Biden moved with his family to Delaware in 1953. He studied at the University of Delaware before earning his law degree from Syracuse University. He was elected to the New Castle County Council in 1970 and to the U.S. Senate in 1972. As a senator, Biden drafted and led the effort to pass the Violent Crime Control and Law Enforcement Act and the Violence Against Women Act. He also oversaw six U.S. Supreme Court confirmation hearings, including the contentious hearings for Robert Bork and Clarence Thomas. Biden ran unsuccessfully for the Democratic presidential nomination in 1988 and 2008. In 2008, Obama chose Biden as his running mate, and Biden was a close counselor to Obama during his two terms as vice president. In the 2020 presidential election, Biden and his running mate, Kamala Harris, defeated incumbents Donald Trump and Mike Pence. Biden is the second Catholic president in U.S. history (after John F. Kennedy), and his politics have been widely described as profoundly influenced by Catholic social teaching.

Taking office at age 78, Biden is the oldest president in U.S. history, the first to have a female vice president, and the first from Delaware. In 2021, he signed a bipartisan infrastructure bill, as well as a $1.9 trillion economic stimulus package in response to the COVID-19 pandemic and its related recession. Biden proposed the Build Back Better Act, which failed in Congress, but aspects of which were incorporated into the Inflation Reduction Act that was signed into law in 2022. Biden also signed the bipartisan CHIPS and Science Act, which focused on manufacturing, appointed Ketanji Brown Jackson to the Supreme Court and worked with congressional Republicans to prevent a first-ever national default by negotiating a deal to raise the debt ceiling. In foreign policy, Biden restored America's membership in the Paris Agreement. He oversaw the complete withdrawal of U.S. troops from Afghanistan that ended the war in Afghanistan, during which the Afghan government collapsed and the Taliban seized control. Biden has responded to the Russian invasion of Ukraine by imposing sanctions on Russia and authorizing civilian and military aid to Ukraine. During the 2023 Israel–Hamas war, Biden announced American military support for Israel, and condemned the actions of Hamas and other Palestinian militants as terrorism.[1] In April 2023, he announced his candidacy for the Democratic Party nomination in the 2024 presidential election.

Early life (1942–1965)

Main article: Early life and career of Joe Biden

Joseph Robinette Biden Jr. was born on November 20, 1942,[2] at St. Mary's Hospital in Scranton, Pennsylvania,[3] to Catherine Eugenia "Jean" Biden (née Finnegan) and Joseph Robinette Biden Sr.[4][5] The oldest child in a Catholic family of largely Irish descent, he has a sister, Valerie, and two brothers, Francis and James.[6]

Joe Biden

Official portrait, 2021

46th President of the United States

Incumbent

Assumed office January 20, 2021

Vice President Kamala Harris

Preceded by Donald Trump

47th Vice President of the United States

In office January 20, 2009 – January 20, 2017

President Barack Obama

Preceded by Dick Cheney

Succeeded by Mike Pence

United States Senator from Delaware

In office January 3, 1973 – January 15, 2009

Preceded by J. Caleb Boggs

Succeeded by Ted Kaufman

Member of the New Castle County Council from the 4th district

In office January 5, 1971 – January 3, 1973

Preceded by Lawrence T. Messick

The text used in the president data set

The different sets are

- **Biden**: https://en.wikipedia.org/wiki/Joe_Biden
- **Trump**: https://en.wikipedia.org/wiki/Donald_Trump
- **Obama**: https://en.wikipedia.org/wiki/Barack_Obama
- **Bush Jr**: https://en.wikipedia.org/wiki/George_W._Bush
- **Clinton**: https://en.wikipedia.org/wiki/Bill_Clinton
- **Bush Sr**: https://en.wikipedia.org/wiki/Bill_Clinton
- **BidenTrump**: Each other paragraph from the Biden and Trump documents

- **Obama to Clinton**: Changed all names from Obama's to Clinton's
    - Barack → Bill, Hussein → Jefferson, Obama → Clinton
- **Biden with noise**: The opening paragraph of the Trump document appended to the Biden document

Sms-spam-detection

Found here: https://archive.ics.uci.edu/dataset/228/sms+spam+collection

Each SMS is put into a separate document and then inputted to the algorithm. The splittxt.py file can be used to split them, since the download is just a large .txt file.

Here, we find some similar or even equal data points. This can be used for spam detection.

Sports articles

https://archive.ics.uci.edu/dataset/450/sports+articles+for+objectivity+analysis

Not really used, since all documents were very different and hence not interesting. Shows a weakness in the.

Movies

https://imsdb.com/

The movie scripts for the following movies:

- Star Wars Episode III
- Star Wars Episode IV
- Star Wars Episode V
- Star Wars Episode VI
- Start Trek
- Lord of the Rings 1
- Lord of the Rings 2
- Lord of the Rings 3