**Practical lessons: 03**

## For python basics: https://cs231n.github.io/python-numpy-tutorial/

**Parsing documents (previous class)**

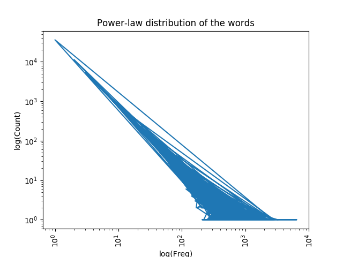
|  |
| --- |
| python parsing\_topics\_documents.py -corpus-path data/WT10G\_LMDIR\_TOP10 -topics-path data/topics451-550.txt.csv -topics-documents-path output/topics\_documents |

**Indexing documents (previous class)**

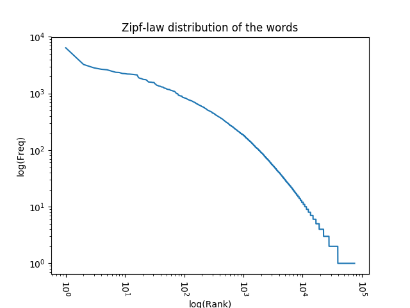
|  |
| --- |
| python indexing\_documents.py -topics-path data/topics451-550.txt.csv -topics-documents-path output/topics\_documents -documents-metadata-path output/documents\_metadata -documents-index-path output/terms\_postinglist |

**Power-law and Zipf-law distribution of the terms**

|  |
| --- |
| python terms\_to\_power\_law\_distribution.py -documents-index-path output/terms\_postinglist |

****

|  |
| --- |
| python terms\_to\_zipf\_law\_distribution.py -documents-index-path output/terms\_postinglist |

****

**Extracting features**

1. **Load** documents metadata from a file
2. **Load** the **Inverted-index** from a file
3. **Extract** topics’ **features**
   1. QL (query length)
   2. **DF** (document frequency)
   3. **IDF** (Inverse document frequency)
4. **Save** the topics and the extracted features to a file

*You can run the following command with the necessary parameters:*

|  |
| --- |
| python feature\_extraction.py -topics-path data/topics451-550.txt.csv -documents-metadata-path output/documents\_metadata -documents-index-path output/terms\_postinglist -topics-features-path output/topics\_features |

NB: Open the output file in “output/topics\_features.” See the format of the data.

**Exercises:**

1. Extract the following features (summation variants) for each query
   1. Sum DF (Summation of document frequency over query terms)
   2. Sum IDF (Summation of inverse document frequency over query terms)
   3. Sum TF (Summation of Term frequency)
   4. Sum TF-IDF (Summation of Term frequency and inverse document frequency)

|  |
| --- |
| python query\_feature\_extraction\_sum\_variants.py -topics-path data/topics451-550.txt.csv -documents-metadata-path output/documents\_metadata -documents-index-path output/terms\_postinglist -topics-features-path output/topics\_features\_sum |

1. Extract the following features (average variants) for each query
   1. Mean DF (Mean of document frequency over query terms)
   2. Mean IDF (Mean of inverse document frequency over query terms)
   3. Mean TF (Mean of Term frequency)
   4. Mean TF-IDF (Mean of Term frequency and inverse document frequency)

|  |
| --- |
| python query\_feature\_extraction\_sum\_mean\_variants.py -topics-path data/topics451-550.txt.csv -documents-metadata-path output/documents\_metadata -documents-index-path output/terms\_postinglist -topics-features-path output/topics\_features\_sum\_mean |

1. Extract the following features (standard deviation variants) for each query
   1. Std DF (Standard deviation of document frequency over query terms)
   2. Std IDF (Standard deviation of inverse document frequency over query terms)
   3. Std TF (Standard deviation of Term frequency)
   4. Std TF-IDF (Std. of Term frequency and inverse document frequency)

|  |
| --- |
| python query\_feature\_extraction\_sum\_mean\_std\_variants.py -topics-path data/topics451-550.txt.csv -documents-metadata-path output/documents\_metadata -documents-index-path output/terms\_postinglist -topics-features-path output/topics\_features\_sum\_mean\_std |

1. Extract the following features (maximum variants) for each query
   1. max DF (Maximum of document frequency over query terms)
   2. max IDF (Maximum of inverse document frequency over query terms)
   3. max TF (Maximum of Term frequency)
   4. max TF-IDF (Maximum of Term frequency and inverse document frequency)

|  |
| --- |
| python query\_feature\_extraction\_sum\_mean\_std\_max\_variants.py -topics-path data/topics451-550.txt.csv -documents-metadata-path output/documents\_metadata -documents-index-path output/terms\_postinglist -topics-features-path output/topics\_features\_sum\_mean\_std\_max |

1. Extract the following features (variance variants) for each query
   1. var DF (Variance of document frequency over query terms)
   2. var IDF (Variance of inverse document frequency over query terms)
   3. var TF (Variance of Term frequency)
   4. var TF-IDF (Variance of Term frequency and inverse document frequency)

|  |
| --- |
| python query\_feature\_extraction\_sum\_mean\_std\_max\_var\_variants.py -topics-path data/topics451-550.txt.csv -documents-metadata-path output/documents\_metadata -documents-index-path output/terms\_postinglist -topics-features-path output/topics\_features\_sum\_mean\_std\_max\_var |