SOEN 479 – PROJECT 1

SPIMI – Indexer
Darrel-Day Guerrero ID:27352409

COMPRESSION TECHNIQUES / NORMALIZING

- **BeautifulSoup4:** Used for extracting content from Reuters21578
- **NLTK:** Used for tokenizing, retrieving English language stop words
- **Compression / Normalizing:**
  - **Case folding**
  - **No numbers (type integer or float removal)**
  - **No punctuation**
  - **No blank or empty strings**
  - **30 stop word removal**
  - **150 stop word removal**
  - ~~**Stemming (e.g. Porter)**~~

COMMANDS

- Clear blocks (/index_blocks) and existing index (spimi_inverted_index.txt)

```
C:\Python36-32\python.exe clear_spimi_files.py
```

- Build the SPIMI inverted index

```
C:\Python36-32\python.exe generate_index.py [block_size_limit]
[block_size_limit] = 75000 default size
```

- Perform document retrieval

```
C:\Python36-32\python.exe query_doc.py
```

QUERIES

- Single word query:

  ```
  >> audi
  Result: [2362, 6481, 11026, 17474, 17570, 19436]
  ```

- Multiple word query (AND):

  ```
  >> nyse && stock && rise
  Result: [205, 19207]
  ```

- Multiple word query (OR):

  ```
  >> turbo || engine
  Result: [264, 344, 627, 652, 727, 933, 970, 1132, 1252, 1497, 1685, 1864, 2718, 2735, 3069, 3185, 3202, 3350, 3398, 3457, 4406, 4568, 4724, 4755,
  5022, 5521, 5825, 6318, 6462, 6535, 6811, 7671, 8365, 8404, 8716, 9353, 9403, 9458, 9662, 9771, 10502, 10740, 10747, 11247, 11404, 12034, 12460,
  12816, 12894, 13049, 13772, 13773, 13774, 13865, 13966, 14031, 14314, 14402, 14748, 14878, 14961, 15014, 15263, 15286, 15314, 15337, 15501,
  15698, 15764, 16557, 16618, 16659, 16691, 17403, 17696, 17797, 17799, 18288, 18659, 19289, 19436, 19637, 20764, 21507, 21523]
  ```