# Assignment 1

## Objective

The objective of this assignment is to solve two problems. The first one is to create tokens for the Crain field database. The next one is to apply stemming to all of those tokens.

## Solution

Below are the decisions regarding the token formation as per the requirements:

1. **How long the program took to acquire the text characteristics?**

The time taken to parse all the files and refine the token is shown in the output and is approximately 1.5 on an average.

1. **How the program handles:**
   1. **Upper and lower case words:**  All tokens are stored in lower case.
   2. **Words with dashes:**  Such words are split into two separate tokens.
   3. **Possessives (e.g. "sheriff's", "university's"):**  “*‘s*” is removed from such tokens.
   4. **Acronyms (e.g., "U.S.", "U.N."):**  The dots are removed from such tokens.
2. **Major Data Structures & Algorithms:** 
   1. **HashMap:** The tokens and its frequency is stored in the hash map with tokens as keys and its frequency as objects.
   2. **Porter’s Algorithm:** For stemming of tokens I have used the implementation of Porter’s Algorithm as provided in the problem statement. [1]
   3. **Jsoup Library:**  I have used this Open Source library to parse the files and remove the SGML tags from them. [2]

## Output

Number of tokens : 237868

Number of unique tokens : 11043

Number of tokens that occur only once : 4831

Average number of tokens per document : 169

30 most frequent words :

Token: "the" freq: 19450

Token: "of" freq: 12717

Token: "and" freq: 6677

Token: "a" freq: 5977

Token: "in" freq: 4645

Token: "to" freq: 4563

Token: "is" freq: 4114

Token: "for" freq: 3493

Token: "are" freq: 2429

Token: "with" freq: 2265

Token: "on" freq: 1944

Token: "flow" freq: 1848

Token: "at" freq: 1834

Token: "by" freq: 1755

Token: "that" freq: 1569

Token: "an" freq: 1388

Token: "be" freq: 1272

Token: "pressure" freq: 1207

Token: "boundary" freq: 1156

Token: "from" freq: 1116

Token: "as" freq: 1113

Token: "this" freq: 1081

Token: "layer" freq: 1002

Token: "which" freq: 975

Token: "number" freq: 972

Token: "results" freq: 885

Token: "it" freq: 857

Token: "mach" freq: 824

Token: "theory" freq: 788

Token: "shock" freq: 712

Time taken is : 1.408 secs

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Number of stems : 237868

Number of distinct stems : 8327

Number of stems that occur only once : 3757

Average number of stems per document : 169

30 Most frequent stems :

Stem: "the" freq: 19450

Stem: "of" freq: 12717

Stem: "and" freq: 6677

Stem: "a" freq: 5977

Stem: "in" freq: 4645

Stem: "to" freq: 4563

Stem: "is" freq: 4114

Stem: "for" freq: 3493

Stem: "ar" freq: 2458

Stem: "with" freq: 2265

Stem: "on" freq: 2262

Stem: "flow" freq: 2079

Stem: "at" freq: 1834

Stem: "by" freq: 1755

Stem: "that" freq: 1569

Stem: "an" freq: 1388

Stem: "pressur" freq: 1382

Stem: "be" freq: 1369

Stem: "number" freq: 1346

Stem: "boundari" freq: 1185

Stem: "layer" freq: 1134

Stem: "from" freq: 1116

Stem: "as" freq: 1113

Stem: "result" freq: 1087

Stem: "thi" freq: 1081

Stem: "it" freq: 1044

Stem: "effect" freq: 995

Stem: "which" freq: 975

Stem: "method" freq: 887

Stem: "theori" freq: 881

## References

1. <http://chianti.ucsd.edu/svn/csplugins/trunk/soc/layla/WordCloudPlugin/trunk/WordCloud/src/cytoscape/csplugins/wordcloud/Stemmer.java>
2. <http://jsoup.org/>