# SPELL CORRECT AND AUTO-COMPLETE

# Spell Check: Peter Norvig Algorithm

Preparation: Big.txt, php code

# 1. big.txt

Parsed through all the downloaded files and created big.txt taking all the data from files for HTML, PDF and Microsoft files, parsed content and created big.txt containing all the textual data from the downloaded data. Fetched all the words respective to the school to provide spell correct for dornsife school.

Created Maven Based Java Application, Used Apache Tika to parse through the files. Configurations: As the crawled data had 50 MB sized crossing TIKA parser limit, parsed half of these files, kept an upper bound of 1000M for big.txt. Parsed content, removed an array of stop words was removed.

```
String[] stopWordsArray = { "i", "me", "my", "myself", "we", "our", "ours",
       "ourselves", "you", "your", "yours", "yourself", "yourselves",
       "he", "him", "his", "himself", "she", "her", "hers", "herself",
       "it", "its", "itself", "they", "them", "their", "theirs",
       "themselves", "what", "which", "who", "whom", "this", "that",
       "these", "those", "am", "is", "are", "was", "were", "be", "been",
       "being", "have", "has", "had", "having", "do", "does", "did",
       "doing", "a", "an", "the", "and", "but", "if", "or", "because",
       "as", "until", "while", "of", "at", "by", "for", "with", "about",
       "against", "between", "into", "through", "during", "before",
       "after", "above", "below", "to", "from", "up", "down", "in", "out",
       "on", "off", "over", "under", "again", "further", "then", "once",
      "here", "there", "when", "where", "why", "how", "all", "any",
       "both", "each", "few", "more", "most", "other", "some", "such",
       "no", "nor", "not", "only", "own", "same", "so", "than", "too",
       "very", "s", "t", "can", "will", "just", "don", "should", "now", "" };
Set<String> stopWordSet = new HashSet<String>/
```

## 2. Configure PHP code

Increase memory limit to 4 MB, and null check in train\_words function, used edit distance as 2 like google does, thus obtained the words like and picks the word with more frequency thus progresve changed to progress.

## Auto Complete: Solr Configurations

1. Create Request Handler:

Configure solrconfig.xml as following:

2. Create Suggest Component

#### Configure solrconfig.xml as following:

Configured build on startup as false to avoid core taking too long to load. Used fuzzy logic look up

```
<searchComponent name="suggest" class="solr.SuggestComponent">
    <!st name="suggester">
        <str name="name">mySuggester</str>
        <str name="lookupImpl">FuzzyLookupFactory</str>
        <!-- org.apache.solr.spelling.suggest.fst -->

        <str name="field">_text_</str>
        <str name="field">_text_</str>
        <str name="suggestAnalyzerFieldType">textSuggest</str>
        <str name="buildOnStartup">false</str>
        <str name="buildOnCommit">false</str>
        </lst>
        </searchComponent>
```

3. Create Custom Suggest Type Field.

Perform tokenization, stemming and create phrase query., used en splitting

Tokenizer used:

```
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
```

Filters used:

Stemmer used:

Porter Stemmer Used from Solr: to convert all the terms to root

<filter class="solr.PorterStemFilterFactory"/>

Additional parameters: To remove duplicates and keywords from the auto spell

<filter class="solr.RemoveDuplicatesTokenFilterFactory"/>

<filter class="solr.KeywordRepeatFilterFactory"/>

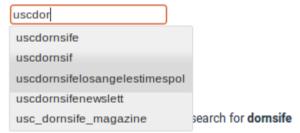
## Build the Component:

The component is built explicitly after the core has been successfully loaded by calling suggest.build=true.

Did not cover N-gram construction to create Autocomplete for multiple terms.

It did not tokenise and showed the text joint like: thus it did predict but not with spaces, If this could be considered although needs improvising.

#### Search:



## Integration:

## 1. Did you mean?

Call Spell Corrector.php and fetch the correct word for each word entered and display the concatenated the result for each corrected word. The phrase links to the hyperlink which fetches the results for corrected phrase.

### 2. Auto Complete:

JSONP call to fetch the results from port 8983, SOLR . Suggest Handler is called and array of phrases is fetched in dropdown. On select of dropdown element corresponding results are displayed.

#### Demo FeedBack:

Home page was coming at 3-4 position as the pages are sorted with score in regular where as when sorted with pagerank homepage comes on top.

Search: Sort by Page Rank Submit Query

#### Search Results

2659 results were found for the search for dornsife

#### Results 1 - 10 of 2659:

#### USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/
File Type: text/html File Size: 24.26KB
Description: The largest and oldest of the USC schools, USC Domsife functions as the academic core of the university, offering courses and advancing knowledge across the humanities, social sciences and sciences.

Author:
Found Keywords: usc, college, usc, college, education, students, graduate, undergraduate, admissions, humanities, social sciences, natural sciences
URL: No URL Found Date: No date found

#### Home > USC Dornsife Gateway Internship Program > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/gatewayInternships
File Type: text/html File Size: 7.78KB
Description: No description Author: No author Found Keywords: usc, college
URL: No URL Found Date: No date found

#### Home > Career Pathways > USC Dana and David Dornsife College of Letters, Arts and Sciences

Link:http://dornsife.usc.edu/careerpathways
File Type: text/html File Size: 8.92KB
Description: USC Dornisfe Student Special Services Career Pathways Website Scholarships, student special services
URL: No URL Found Date: No date found

Keywords: usc, college, career, co-curricular, curricular, employment, fellowships, graduate school, pathways, professions that she was a constant of the constant

Thus relevance of home page became less compared to other publications page because of the lucene search.