CPEG 657 - Search & Data Mining 2/22

Picking the right tool

```
* Is it a retrieval related task?
* A machine learning related tasks?
* Other Tasks?
```

Search Related Tasks

- Lemur
  1. C/C++ based
  2. Indri - langauge model based retrieval platform | https://www.lemurproject.org/indri.php
     Good for natural language - maybe?
  3. Lemur - a more general platfrom for most retrieval models | https://www.lemurproject.org/lemur.php
     Good for building keyword based search engine

- Solr
  1. http://www.apache.org/dyn/closer.lua/lucene/solr/6.4.1 | https://wiki.apache.org/solr/SolPython
  2. Java based
  3. A powerful search engine
  4. Provide search in field functionality to my search engine
  5. Has user interface that is customizable
  6. Difficult to modify

Machine Learning Related Projects

- Weka
  1. http://www.cs.waikato.ac.nz/ml/weka/ | https://pypi.python.org/pypi/python-weka-wrapper | https://github.com/fracpete/python-weka-wrapper
  2. GUI Java or command line
  3. good for prediction, classify, or develop new machine learning algorithm
  4. Building the training/testing set
  5. Features engineering
  6. Compare different algorithms

Text Summarization

- MEAD Summarization
    1. http://www.summarization.com/mead/
    2.

Java

- Anserini
    1. Indexing and Searching

- Galago
    1. Indexing and Searching

- Stanford Core NLP
    1. NLP
    2. Parts of Speech
    3. Sentiment Analysis

- Ranklib
    1. Machine learning with ranking

- Anserini
    1. Indexing and Searching

Python

- NLTK
    1. http://www.nltk.org
    2. Similar to Stanford Core NLP

- Scikit-learn
    1. Machine Learning
    2. Classification, Regression, DT
    3. Topic Modeling
    4. USE THIS!!!!!

- TensorFlow
    1. Deep Learning
    2. LSTM
    3. Word2Vec

VIRLab (1st Assigment) | http://infolab.ece.udel.edu:8008

- Web-based virtual lab for IR
- Implementation
    1. score - ranking score of document
    2. Document Frequency (DF)
    3. Be careful with math