Information Retrieval: Assignment 2

Kirill Semenov
NPFL103, 3rd January 2023

Chosen Package

Terrier

- Java-based, but has Python wrapper
- Big functionality in preprocessing
  - But quite inconvenient for Python
- Big functionality in indexing/query expansion/etc.
  - Very convenient in Python
- Problematic to run on Windows
  - I ran on colab

Implementation:

- Preprocessing: my own
- Indexing, retrieval, query expansion: PyTerrier

Experiments

18 experiments

Parameters:

1. Preprocessing
2. Tokenization (+ lemmatization + stopwords)
3. Indexing functions
4. Enhanced query construction
5. Query expansion

Preprocessing & Tokenization

1. Case: sensitive VS insensitive
2. Tokenization: nltk VS Morphodita
3. Stopwords: on VS off

Run/experiment	**MAP (EN)**	**MAP (CS)**
Run-0	0.2678	0.2056
Case-insensitive	**0.3602**	**0.2542**
Morphodita	0.1047	0.0728
Nltk	0.3602	0.2542
Clear-stopwords	0.3588	0.2518

Indexing, Weighting Models

TF, BM25, TF-IDF - different versions of vector space models

PL2, LGD - Divergence-from-randomness models

Dirichlet LM, LGD, Hiemstra LM - language LMs

Run/ experiment	**MAP (EN)**	**MAP (CS)**
BM25	0.3497	0.2517
TF	0.1234	0.0605
PL2	0.3355	0.2405
Dirichlet LM	0.3498	0.2654
Run 1: LGD	**0.3887**	**0.2735**
NLTK+LGD	**0.3887**	**0.2735**
Hiemstra LM	0.3018	0.2339

Query Construction & Expansion

Query construction: Title VS Title + Description

Query Expansion: Bo1

Run/experiment	**MAP (EN)**	**MAP (CS)**
Tf-idf + query expansion	0.3669	0.3102
Bm25 + query expansion	0.3613	0.3084
Run 2: LGD + query expansion	**0.3686**	**0.3111**
DPH + query expansion	0.3626	0.3033
LGD + enhanced	0.3084	0.2488
BM25 + enhanced	0.2945	0.2488

Results

1. Case insensitivity - good
2. LMs + Divergence from randomness - good
3. Query expansion - good (especially for Czech)
4. Run-1: LGD + case insensitivity (best English performance)
5. Run-2: LGD + case insensitivity + query expansion (best Czech performance)
6. Things to explore more in PyTerrier:
    a. Other options of query expansion
    b. Neural IR

Thank you for your attention!