Information Retrieval

Assignment 4:
Synonym Expansion with Lucene and WordNet

Patrick Schäfer (patrick.schaefer@hu-berlin.de)

Marc Bux (buxmarcn@informatik.hu-berlin.de)

Synonym Expansion

- Idea: When a user searches a term K, implicitly search for all synonyms of K
  - S AND T -> (S OR S' OR ..) AND (T OR T' …)
- Popular method
- Usually increases recall and decreases precision
- Requires a high quality synonym lexicon
- Can be extended to also include hyponyms ('banana' is a hyponym to 'fruits').

WordNet

- Lexical database with semantic relationships
- Maintained since 1985
- Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets).
- ~60.000 words, ~160.000 Synsets

- Contains different relationship types: hypernomy, hyponomy, causation, antonomy, holonym, meronym …

Some Relationship Types

- Antonyms are words with opposite meanings:
  bad is an antonym of good
- Hyponyms are specific instances of a category:
  red is a hyponym of color
- Hypernyms describe categories of instances:
  color is a hypernym of red
- Holonyms define a relationship between terms (one is part of the other):
  tree is a holonym of trunk
- Meronyms are the opposite of holonyms:
  trunk is a meronym of tree

Task

- Implement synonym expansion within Lucene (v6.3) for the IMDB movie plots.
- You can reuse your existing code from assignment 3 (using word tokenization and stop word removal, but no stemming).
- Use WordNet as lexicon
  - current release, WordNet 3.1
- For simplicity, we will only consider Boolean (AND, OR, NOT) term search.
- No phrase or proximity search any more

Example Synsets from WordNet

[well]: [considerably] [intimately] [easily] [comfortably] [wellspring] [substantially] [advantageously] [good] [swell] [fountainhead]

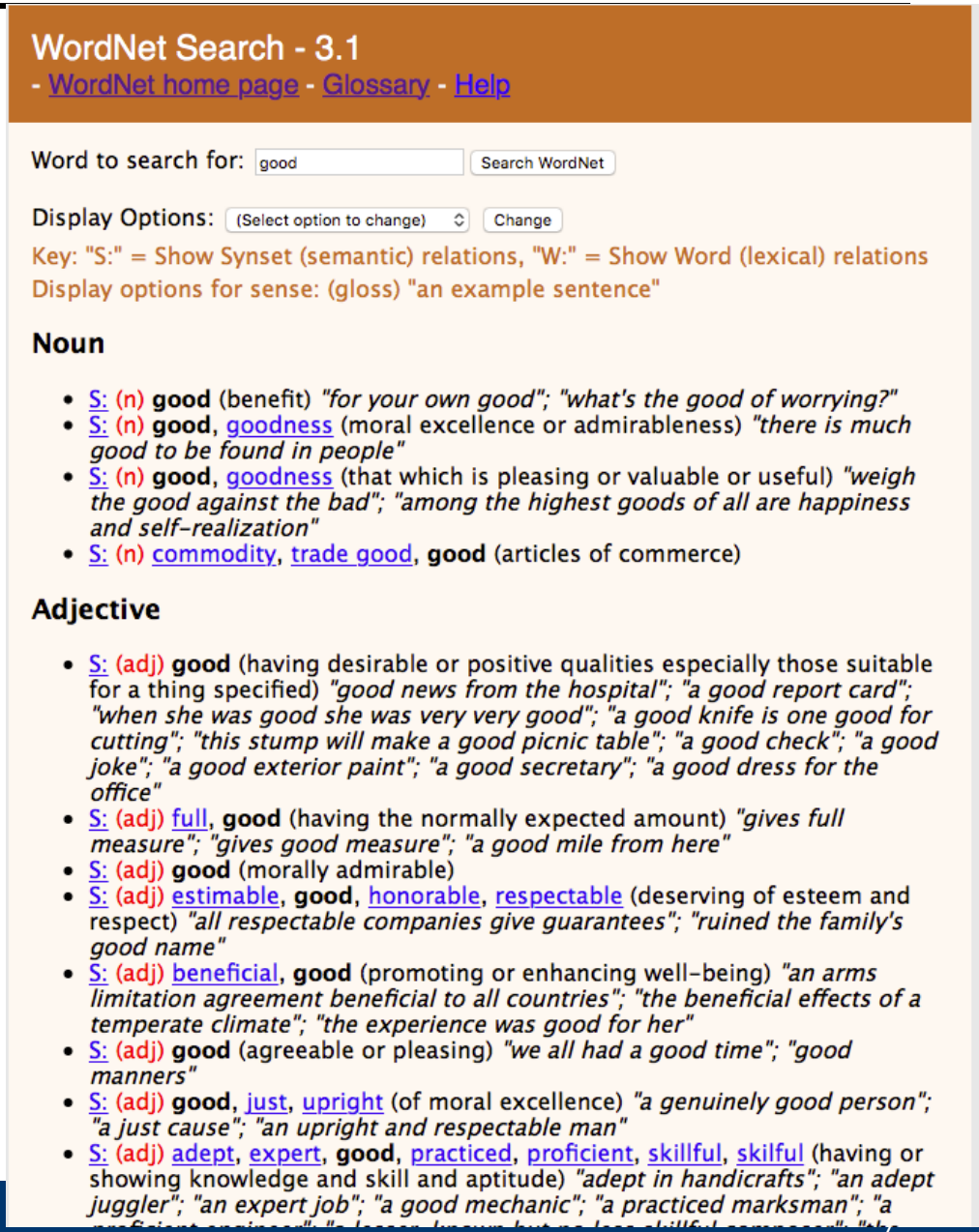
[good]: [commodity] [expert] [sound] [respectable] [secure] [estimable] [effective] [honest] [serious] [ripe] [near] [unspoiled] [dear] [just] [salutary] [goodness] [proficient] [skilful] [adept] [thoroughly] [soundly] [unspoilt] [dependable] [right] [upright] [beneficial] [safe] [well] [honorable] [full] [practiced] [skillful]

[better]: [expert] [sound] [secure] [good] [wagerer] [honest] [serious] [easily] [near] [unspoiled] [salutary] [goodness] [adept] [bettor] [amend] [soundly] [intimately] [dependable] [comfortably] [upright] [improve] [beneficial] [safe] [punter] [wellspring] [substantially] [advantageously] [full] [skillful] [commodity] [meliorate] [respectable] [best] [swell] [fountainhead] [estimable] [effective] [ripe] [dear] [just] [proficient] [skilful] [thoroughly] [break] [considerably] [unspoilt] [right] [ameliorate] [well] [honorable] [practiced]

Wordnet

- You can search synsets directly at WordNet:

  http://wordnetweb.princeton.edu/perl/webwn



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: good Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **good** (benefit) *"for your own good"; "what's the good of worrying?"*
- S: (n) **good**, goodness (moral excellence or admirableness) *"there is much good to be found in people"*
- S: (n) **good**, goodness (that which is pleasing or valuable or useful) *"weigh the good against the bad"; "among the highest goods of all are happiness and self-realization"*
- S: (n) commodity, trade good, **good** (articles of commerce)

Adjective

- S: (adj) **good** (having desirable or positive qualities especially those suitable for a thing specified) *"good news from the hospital"; "a good report card"; "when she was good she was very very good"; "a good knife is one good for cutting"; "this stump will make a good picnic table"; "a good check"; "a good joke"; "a good exterior paint"; "a good secretary"; "a good dress for the office"*
- S: (adj) full, **good** (having the normally expected amount) *"gives full measure"; "gives good measure"; "a good mile from here"*
- S: (adj) **good** (morally admirable)
- S: (adj) estimable, **good**, honorable, respectable (deserving of esteem and respect) *"all respectable companies give guarantees"; "ruined the family's good name"*
- S: (adj) beneficial, **good** (promoting or enhancing well-being) *"an arms limitation agreement beneficial to all countries"; "the beneficial effects of a temperate climate"; "the experience was good for her"*
- S: (adj) **good** (agreeable or pleasing) *"we all had a good time"; "good manners"*
- S: (adj) **good**, just, upright (of moral excellence) *"a genuinely good person"; "a just cause"; "an upright and respectable man"*
- S: (adj) adept, expert, **good**, practiced, proficient, skillful, skilful (having or showing knowledge and skill and aptitude) *"adept in handicrafts"; "an adept juggler"; "an expert job"; "a good mechanic"; "a practiced marksman"; "a*

Query Expansion in Lucene

- There are two options:
- At indexing time: Add all expansions to all terms of a document d when indexing d.
- At search time: When searching a keyword K, rewrite query in disjunction of all expansions of K.
  - Query: plot:Berlin AND plot:wall AND type:television
    - plot:berlin AND (plot:bulwark OR plot:fence OR plot:palisade OR plot:paries OR plot:rampart OR plot:surround OR plot:wall) AND (type:telecasting OR type:television OR type:telly OR type:tv OR type:video)

- Note: If K is part of more than one synset, use all
  - No disambiguation

Getting Started

- Download WordNet 3.1 files at
  - http://wordnetcode.princeton.edu/wn3.1.dict.tar.gz

- Extract noun, verb, adj, adv files:
  - data.[noun, verb, adj, adv] (synsets)
  - [noun, verb, adj, adv].exc (base forms)

- Parse synsets from these plain files using syntax:
  - http://wordnet.princeton.edu/man/wndb.5WN.html

Data File Format

- Each data file begins with a copyright notice. Skip this.
- Each synset is encoded in one line.
- Each line has the format:

  synset_offset lex_filenum ss_type w_cnt word lex_id [word lex_id...]
  p_cnt [ptr...] [frames...] | gloss
- *w_cnt*: Two digit hexadecimal integer indicating the number of words.

- Example line (synset):

  00007846 03 n 06 person 0 individual 0 someone 0 somebody 0 mortal 0 soul 0 421 @ 00004475 n 0000 @ 00007347 n 0000 #m 07958392 n 0000 + 01562007 a 0501 + 00388736 v 0203 + 04626138 n 0101 + 00729535 v 0101 %p 04624919 n 0000 %p ...

Exception List File Format

- The first field of each line is an inflected form, followed by a space separated list of one or more base forms of the word.

- Examples:

  better good well

  bigger big

- Meaning: all synsets of good and well apply to better (but not the reverse).

Complications I

- Use only single-token synonyms
  - Ignore all synonyms with more than one token
  - These are formatted by a "_" in the name (e.g., house_of_cards)

- Special adjective syntax
  - Remove (p), (a) and (ip) from adjectives (e.g. galore(ip)).
  - https://wordnet.princeton.edu/man/wninput.5WN.html

Complications II

- Merge synsets of words appearing in the verb, nouns, adj, adv files, such as reason (noun) and reason (verb).

- Consider a synset as set
  - Example: Synset of cause = {reason,grounds}
  - Create the following synonym relations: cause-reason, cause-grounds, reason-grounds and all reverse relations reason-cause, grounds-cause, grounds-reason.

- BUT do not apply this rule transitively
  - Example: cause = {grounds} and grounds={earth} should not create cause-earth!
  - Syn-relationships in WordNet do not form an equivalence class

Complications III

- The exception lists are not symmetric. The inflected form is merged with all synsets of its base forms but not the reverse:
- Exception: better good well
  - syns(better) := syns(better) ∪ syns(good) ∪ syns(well) ∪ good ∪ well
  - But not syns(well) := syns(better) ∪ …
- And do not apply this rule transitively.
- Some true results for reference:

  Only sysnets: 60993 words with 153394 synonyms.

  Synsets & exception lists: 60993 words with 159182 synonyms.

Getting started

- in BooleanSeachWordnet.java, implement the functions:
  - public void buildSynsets(String wordnetDir)
    (used to parse the wordnet files and build the synonym index)
  - public void buildIndices(String plotFile)
    (used to parse the file and build the lucene index)
  - public Set<String> booleanQuery(String queryString)
    (parses the query string and returns the title lines of any entries in the plotFile matching the query)
  - public void close()
    (can be used to close Lucene index, Threadpool, etc.)

Test your Program

- we provide you with a modified:
  - queries_wordnet.txt file containing exemplary queries
  - results_wordnet.txt file containing the expected results of running these queries
  - main method for testing your code (which expects as parameters the corpus file, the queries file and the results file)

- you can check your synonym expansion for plausibility on the WordNet website:
  - http://wordnetweb.princeton.edu/perl/webwn

Deliverables

- by Thursday, 26.01.17, 23:59 (midnight)
- submission: archive (zip, tar.gz)
  - contains Java source files, any used libraries, and your compiled jar named BooleanQueryWordnet.jar
  - file name (of submitted archive): your group name
- upload to https://hu.berlin/24377
  - if this doesn't work, send via mail to buxmarcn@informatik.hu-berlin.de
- test your jar before submitting by running our queries on gruenau2
  - java -jar BooleanQueryWordnet.jar <plot list file> <wordnetDir> <queries file> <results file>
  - you might have to increase the JVM's heap size (e.g., -Xmx8g)
  - your jar must run and answer all test queries in 'queries_wordnet.txt' correctly

Presentation of Solutions

- you are be able to pick when and what you'd like to present (first-come-first-served):
  - monday: https://dudle.inf.tu-dresden.de/inforet_ue4_mo/
  - tuesday: https://dudle.inf.tu-dresden.de/inforet_ue4_tu/

- presentation will be given on 30./31.01.17

- One team can present their Lucene WordNet Indexer.
- Two teams can present their Lucene Query Expansion.

Competition

- Search as fast as possible.
- stay under 40 GB memory usage.
- we will call the program using our eval tool:
  - we will use different queries and -Xmx40g parameter

- We will evaluate twofold:
  a) The total query time.
  b) The total time for building the index.

Checklist

again, before submitting your results, make sure that you

1. did not change or remove any code from BooleanQueryWordnet.java
2. did not alter the functions‘ signatures (types of params, return values)
3. only use the default constructor and don‘t change its parameters
4. did not change the class or package name
5. named your jar BooleanQueryWordnet.jar
6. tested your jar on gruenau2 by running
   java -jar BooleanQueryWordnet.jar plot.list wordNetDir
   queries_wordnet.txt results_wordnet_wordnet.txt
   (you might have to increase Java heap space, e.g. -Xmx6g)
7. ascertained that the queries in queries.txt were answered correctly
8. Make sure to upload a zip file named by your group name.

Next Steps

- this week: evaluation of assignment 3
- next weeks: Q/A sessions for assignment 4.
- Upload your solution by Thursday, 26.01.17, 23:59 (midnight)