Text Processing

Information Retrieval

Inf 141 / CS 121

Tokenization

- Break the input into words
  - Character stream -> token stream
  - Called a tokenizer / lexer / scanner
- Compiler front-end
  - Lexer hooks up to parser
- Preprocessor for information retrieval
  - Lexer feeds tokens to retrieval system

Identifying Tokens

- Divide on whitespace and throw away punctuation?
- What is a token? Depends…
  - Apostrophes
    - O’Neill
    - aren’t
  - Hyphen-handling
    - clear-headed vs clearheaded
    - mother-in-law

Identifying Tokens

- Multiple words as single token?
  - San Francisco
  - white space
  - New York University vs York University
- Tokens that aren't words
  - jossher@uci.edu
  - http://www.ics.uci.edu/~lopes
  - 192.168.0.1

Markup as Tokens

- Many documents are structured using markup
  - HTML, XML, ePub, etc…
- What to do about tags?
  - Include them as tokens
  - Filter them out entirely
  - Filter tokens based on tags

Advanced Tokenization

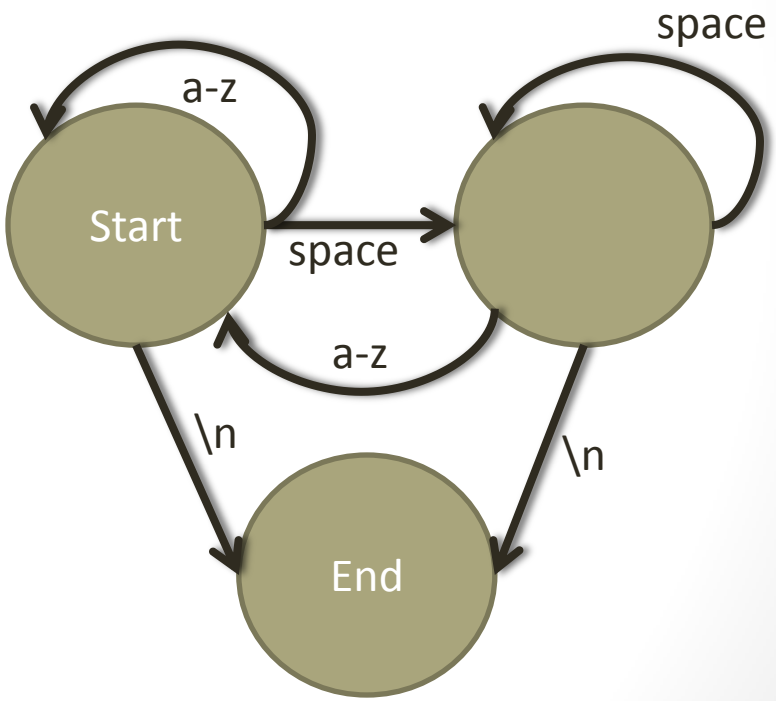
- Tokenization can do more than break a character stream into tokens
- Programming language tokenizers use specific grammars
  - Can identify comments, literals
  - Associate a type with each token

Writing a Tokenizer

- `while` loop looking for delimiters
  - Fast to write and execute
  - Hard to maintain and easy to mess up
- Java library methods
  - `java.util.Scanner`
  - `java.util.String.split()`
  - `java.util.StringTokenizer`

Writing a Tokenizer

- Deterministic Finite Automaton (DFA)
  - Finite set of states
  - Alphabet
  - Transition function
  - Start state
  - End states



Generating a Tokenizer

- Numerous open source tools
  - ANTLR, JFlex, JavaCC
  - Usually focused on programming languages
- Specify the grammar, tool generates the program
  - Easy to maintain
  - Very flexible

Dropping Common Terms

- Very common words can be bad for IR systems
  - he has it on that and as by with a…
- Stop words
  - Use up lots of space in the index
  - Match nearly every document
  - Rarely central to document's meaning
- How to detect them?
  - Assignment part b

Drop Common Terms?

- Should you remove stop words?
  - Flights to London vs Flights London
  - Flights from London vs Flights London
  - How to search for “to be or not to be”?
- Trend in Information Retrieval is to not use stop words
  - Replaced by statistical techniques

Normalization

- Canonicalize tokens so that superficial differences don’t matter
  - USA = U.S.A. = usa
  - C.A.T = cat?
- Techniques
  - Remove accents & diacritics
  - Case-folding
  - Collapse alternate spellings

Stemming and Lemmatisation

- Reduce word variants to single version
  - am, are, is => be
- Stemming
  - Reduce words to *stem* by chopping off suffix
- Lemmatization
  - Remove inflection to arrive at base dictionary form of the word, called a *lemma*

Porter's Algorithm

- Most common algorithm for stemming English
  - 5 phases of sequential word reduction
- Stage 1 example
  - SSES -> SS caresses -> caress
  - IES -> I ponies -> poni
  - SS -> SS caress -> caress
  - S -> cats -> cat

Stemming Example

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Stemming vs Lemmatisation

- Stemming is easy (ish)
  - Fairly simple set of rules
- Lemmatisation is hard
  - Requires complete vocabulary and morphological analysis
- Which is better for retrieval?
  - Depends…
  - Both improve recall and harm precision

Acronym Expansion

- Expands acronyms and abbreviations into their full form
  - USA -> united stats of america
  - In4matx -> informatics
- Usefulness depends on domain
  - Source code retrieval greatly aided

Language Differences

- Some languages have more morphology than English
  - Spanish, German, Latin
- German has compound words
- Chinese and Japanese don't segment words
- French for *the* is a prefix that changes depending on context