Informatics 225
Computer Science 221

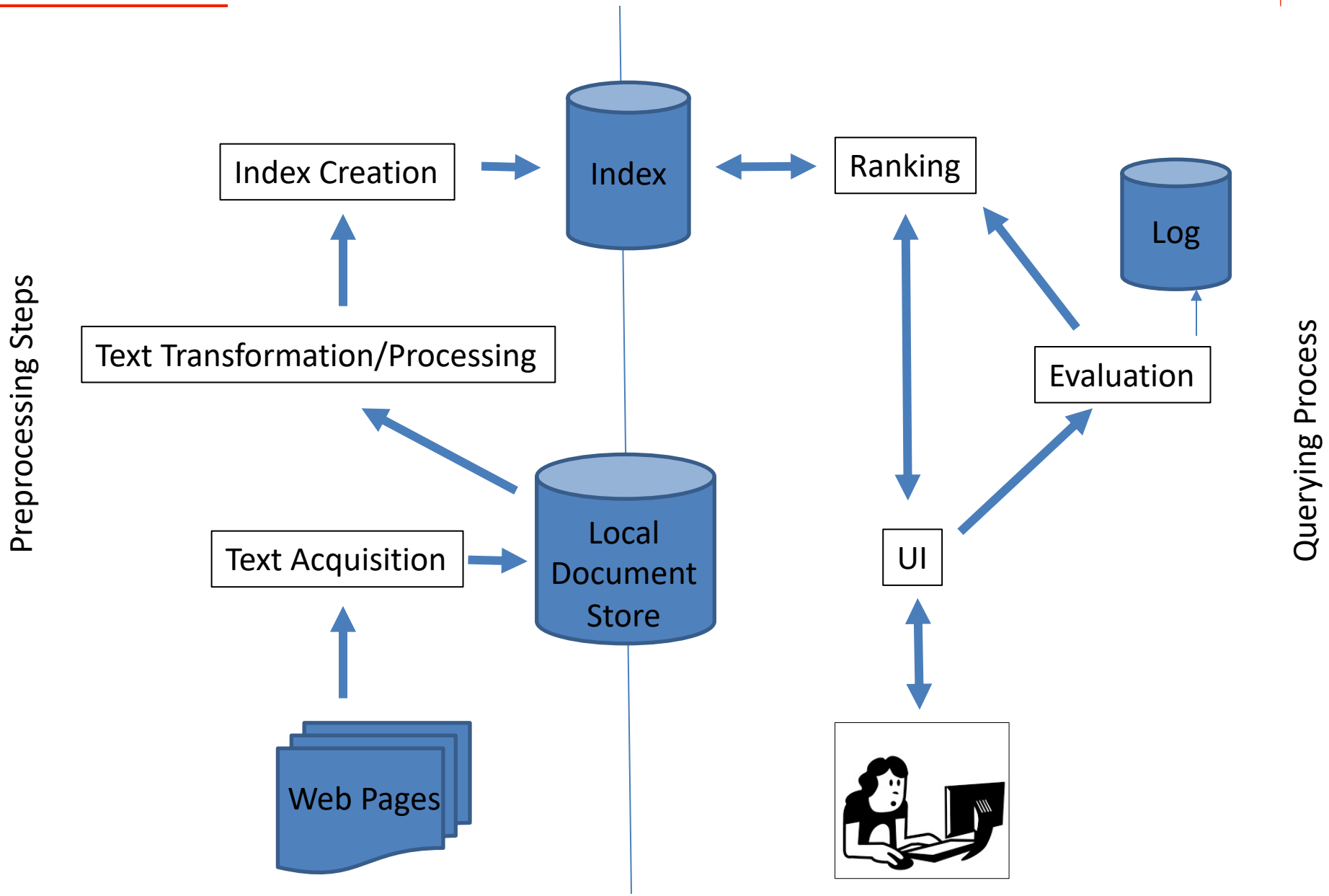
Information Retrieval

Lecture 13

Duplication of course material for any commercial purpose without the explicit written permission of the professor is prohibited.

These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Prof. Alberto Krone-Martins, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Powerpoint theme by Prof. André van der Hoek.

Architecture



Text processing & tokenization++

Information Retrieval

Tokenizing or lexical analysis

- Forming units of meaning from characters (often, sequences)

Tokenizing or lexical analysis

- Forming units of meaning from characters (often, sequences)
- Surprisingly complex in English
  - can be harder in other languages
- Early IR systems (not necessarily web search!):
  - any sequence of alphanumeric characters of length 3 or more
  - terminated by a space or other special character
  - upper-case changed to lower-case *(case folding, normalization or downcasing)*

Tokenizing

- Simple early IR tokenizer :
  - "Bigcorp's 2007 bi-annual report showed profits rose 10%."

  becomes
  - "bigcorp 2007 annual report showed profits rose"

- Too simple for search or even large-scale experiments

- Why? *Too much information lost*
  - Small decisions in tokenizing **can have major impact** on effectiveness of some queries

Tokenizing problems to consider

- Small words can be important in some queries, usually in combinations
  - xp, am, pm, el paso, world war II, “to be or not to be”
- Both hyphenated and non-hyphenated forms of many words are common
  - Sometimes hyphen is not needed *(and you can merge)*
    - e-bay, wal-mart, active-x, cd-rom, t-shirts
  - At other times, hyphens should be considered either as part of the word or a word separator *(and it is harder to decide what to do)*
    - winston-salem, mazda rx-7, e-cards, pre-diabetes, t-mobile, spanish-speaking

Tokenizing problems to consider

- Special characters are an important part of tags, URLs, code in documents

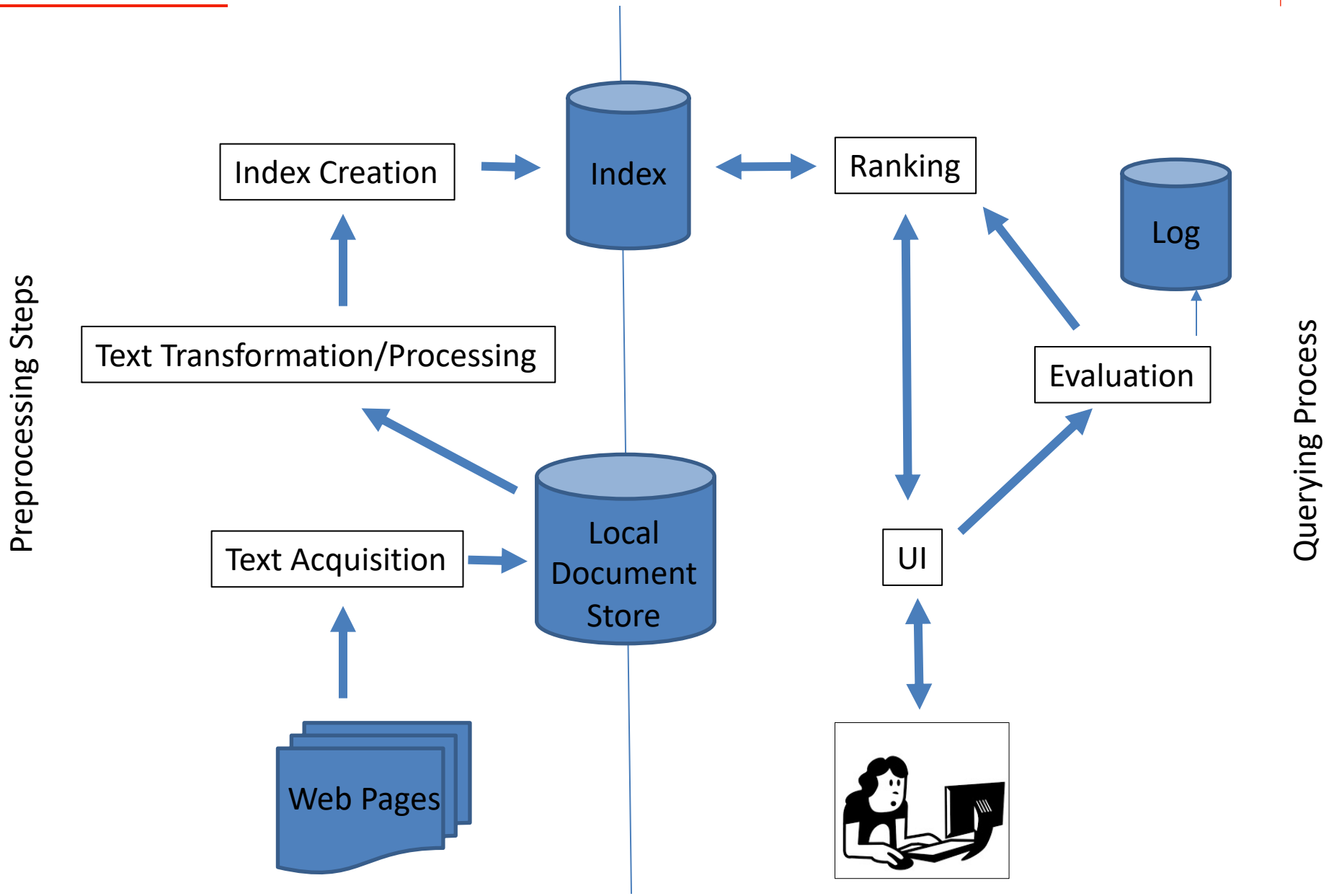
- Capitalized words can have different meaning from lower case words
  - “Bush”, “Apple” may, or may not be the same as “bush”, “apple”

- Apostrophes can be a part of a word, a part of a possessive, or just a mistake
  - rosie o'donnell, can’t, don't, 80's, 1890's, men's straw hats, master's degree, england's ten largest cities, shriner's

Tokenizing problems to consider

- Numbers can be important, including decimals
  - nokia 3250, top 10 courses, united 93, quicktime 6.5 pro, 92.3 the beat, 288358, 3.14159265
- Periods can occur in numbers, abbreviations, URLs, ends of sentences, and other situations
  - I.B.M., Ph.D., cs.umass.edu, F.E.A.R.
- Note: in a search engine, tokenizing steps for queries must be identical to steps for documents

Architecture



Tokenizing Process

- Multiple step process.
- First step is to use tokenizer+parser to identify appropriate parts of document to tokenize for search purposes
- Most general strategy: to defer complex decisions to other components
  - word is any set of alphanumeric characters, terminated by a space or special character, with everything converted to lower-case
  - everything indexed, and the system keeps track of additional information
    - example: 92.3 → 92 3 but search finds documents with 92 and 3 adjacent
  - incorporate some special rules to reduce dependence on query transformation components

Tokenizing Process

- Not very different from simple tokenizing process used in past
- Examples of rules used with early TREC collections
  - Apostrophes in words ignored
    - o'connor → oconnor bob's → bobs
  - Periods in abbreviations ignored
    - I.B.M. → ibm Ph.D. → phd
- But not ignoring post-processing that will perform information extraction and query transformation for difficult cases.

Stopping

- Function words (determiners, prepositions) have little meaning on their own
- High occurrence frequencies
- Can be treated as *stopwords* (i.e. removed)
  - reduce index space, improve response time, improve effectiveness
- However : can be important in combinations
  - e.g., “to be or not to be”

Stopping

- Stopword list can be created from high-frequency words or based on a standard list
- Lists are customized for applications, domains, and even parts of documents
  - e.g., “click” and “here” are good stopwords for anchor text
- Usually: Best policy is to index all words in documents, make decisions about which words to use at query time

Stemming (or conflation)

- Morphemes
  - Stems : core units
  - Affixes : units that adhere to stems

- E.g. :
  - Stems : STEM is a stem, S is the affix

- Many morphological variations of words
  - *inflectional* (plurals, tenses)
  - *derivational* (making verbs nouns by adding suffix-ation etc.)

- In most cases, these have the same or very similar meanings

Stemming (or conflation)

- **Stemmers** attempt to reduce morphological variations of words to a common stem
  - usually involves removing suffixes
- Can be done at indexing time and/or as part of query processing (just like stopwords)
- Example
  - User query: “mark spitz swimming”.
  - In the search engine corpus, the webpages might contain “swam” as they are usually talking about events that already happened.
  - So it is up to the stemmer to transform “swimming” and “swam” into the same stem (e.g. “swim”) enabling the match to be determined.

Stemming (or conflation)

- Generally a small but significant effectiveness improvement
  - can be crucial for some languages
  - e.g., 5-10% improvement for English, up to 50% in Arabic or Russian

ki**t**a**b**	*a book*
ki**t**a**b**i	*my book*
a**lk**i**t**a**b**	*the book*
ki**t**a**b**uki	*your book* (f)
ki**t**a**b**uka	*your book (m)*
ki**t**a**b**uhu	*his book*
ka**t**a**b**a	*to write*
ma**k**t**a**b**a**	*library, bookstore*
ma**k**t**a**b	*office*

Words with the Arabic root ktb

Stemming (or conflation)

- Two basic types
  - Dictionary-based:
  - Algorithmic:

Stemming (or conflation)

- Two basic types
  - Dictionary-based:
    - Uses lists of related words
    - Language dependent
  - Algorithmic:
    - Uses program to determine related words
    - Language dependent

Stemming (or conflation)

- Two basic types
  - Dictionary-based:
    - Uses lists of related words
    - Language dependent
  - Algorithmic:
    - Uses program to determine related words
    - Language dependent

- Algorithmic stemmers
  - Simplest English algorithmic stemmer
    - *suffix-s:* remove 's' endings assuming plural
      - e.g., cats → cat, lakes → lake, wiis → wii
      - Many *false negatives.* Cannot transform: centuries → century
      - Some *false positives.* It will associate is → I

Porter Stemmer

- Algorithmic stemmer used in IR experiments since the 70s
- Consists of a series of rules. At each step, the rule for the longest applicable suffix is executed.
- Effective in TREC
- Produces *stems* not *words*
- Drawback: Makes a number of errors and difficult to modify

Porter Stemmer

- Example step (1 of 5)

Step 1a:

- Replace ***sses*** by ***ss*** (e.g., stresses → stress).
- Delete ***s*** if the preceding word part contains a vowel not immediately before the ***s*** (e.g., gaps → gap but gas → gas).
- Replace ***ied*** or ***ies*** by ***i*** if preceded by more than one letter, otherwise by ***ie*** (e.g., ties → tie, cries → cri).
- If suffix is ***us*** or ***ss*** do nothing (e.g., stress → stress).

Step 1b:

- Replace ***eed***, ***eedly*** by ***ee*** if it is in the part of the word after the first non-vowel following a vowel (e.g., agreed → agree, feed → feed).
- Delete ***ed***, ***edly***, ***ing***, ***ingly*** if the preceding word part contains a vowel, and then if the word ends in ***at***, ***bl***, or ***iz*** add ***e*** (e.g., fished → fish, pirating → pirate), or if the word ends with a double letter that is not ***ll***, ***ss*** or ***zz***, remove the last letter (e.g., falling→ fall, dripping → drip), or if the word is short, add ***e*** (e.g., hoping → hope).
- Whew!

Porter Stemmer

False positives	*False negatives*
organization/organ	european/europe
generalization/generic	cylinder/cylindrical
numerical/numerous	matrices/matrix
policy/police	urgency/urgent
university/universe	create/creation
addition/additive	analysis/analyses
negligible/negligent	useful/usefully
execute/executive	noise/noisy
past/paste	decompose/decomposition
ignore/ignorant	sparse/sparsity
special/specialized	resolve/resolution
head/heading	triangle/triangular

Porter Stemmer

False positives	*False negatives*
organization/organ	european/europe
generalization/generic	cylinder/cylindrical
numerical/numerous	matrices/matrix
policy/police	urgency/urgent
university/universe	create/creation
addition/additive	analysis/analyses
negligible/negligent	useful/usefully
execute/executive	noise/noisy
past/paste	decompose/decomposition
ignore/ignorant	sparse/sparsity
special/specialized	resolve/resolution
head/heading	triangle/triangular

- Porter2 stemmer addresses some of these issues (and you can include exceptions)
  - Porter : https://tartarus.org/martin/PorterStemmer/
  - Porter2 : http://snowball.tartarus.org/algorithms/english/stemmer.html

- Similar approach has been used with other languages

Dictionary stemmers

- Human-built word lists
  - Pairs of <word, stem>
  - Very low false positive rate
  - Related words do not need to be similar : e.g. “is”, “be”, “was”
  - Important drawback : cannot react automatically to new words.
  - We can build lists semi-automatically by statistical analysis, and then perform human curation of the lists
    - This is useful for query expansion in search applications

Krovetz Stemmer

- Typically irregular words are the oldest in the language; new words usually follow more regular grammatical conventions.

- Hybrid algorithmic-dictionary
  - Word checked in dictionary
    - If present, either left alone or replaced with “exception”
    - If not present, word is checked for suffixes that could be removed
    - After removal, dictionary is checked again

Krovetz Stemmer

- Typically irregular words are the oldest in the language; new words usually follow more regular grammatical conventions.
- Hybrid algorithmic-dictionary
  - Word checked in dictionary
    - If present, either left alone or replaced with “exception”
    - If not present, word is checked for suffixes that could be removed
    - After removal, dictionary is checked again
- Advantage: *usually produce words not stems*
- Lower false positive rate, somewhat higher false negative (depending on the size and quality of curation of the dictionary)
- Comparable effectiveness for search applications

Stemmer Comparison

Original text:

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Porter stemmer:

document describ market strategi carri compani agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale market share stimul demand price cut volum sale

Stemmer Comparison

Original text:

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Porter stemmer:

document describ market strategi carri compani agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale market share stimul demand price cut volum sale

Krovetz stemmer:

document describe marketing strategy carry company agriculture chemical report prediction market share chemical report market statistic agrochemic pesticide herbicide fungicide insecticide fertilizer predict sale stimulate demand price cut volume sale