Regular Expressions

School of Information Studies
Syracuse University

Regular Expressions

Notation for specifying patterns of text

Python package to define patterns

Functions to find pattern matches

Examples:

- Names
- Email addresses
- Phone numbers
- URLs
- Dates

Simple Text Matching

Want to count occurrences of words

RegEx	Description	Example of Match
z	Matches any z	Lazy
[wW]	A single w or W	Woodchuck, woodchuck
[0-9]	Matches one of the digits	Chapter 1
[A-Z]	Any capital letter	Pearl Jam
.	Matches any character	Lazy

Regular Expressions in Python

Use pattern = re.compile (“<regular expression>”)

Match function—true result

Function findall ()—list of results

Can use substitution

Text Matching

RegEx	Description
. (period)	Matches any character
^	Means NOT those characters
|	Match alternatives
A?	Previous object is optional
A*	0 or more of previous object
A+	1 or more of previous object
C(he)?at	Matches Cat or Cheat

Anchors

RegEx	Description
^The	Match must occur at beginning of text
End$	Match must occur at end of text
\b	Match must occur at word boundary
\B	Match must occur at not a word boundary

Escapes

RegEx	Description
\.	Matches the character '.'
\n\t	Match newline, tab
\s	Any character of white space
\d	Any digit
\w	Any word character [A-Za-z0-9]
\S	Any character NOT Whitespace
\D	Any character not a digit
\W	Any character not a word character

Sentiment Analysis

School of Information Studies
Syracuse University

Sentiment Analysis

Used in numerous situations

- Reference a person’s attitude
  - Public sentiment in Twitter
  - Gallup polls
- Positive or negative sentiment toward a movie
  - Opinions
- Product reviews
  - Opinions
  - Different aspects of product

Facts—people, places, things, events

Non-factual aspects—affective or subjective

Scherer Typology of Affective States

State	Description
Emotion	Brief, organically synchronized—*angry, sad, joyful, fearful, ashamed, proud, elated*
Mood	Diffuse, non-caused, low-intensity, long-duration change in subjective—*cheerful, gloomy, irritable, listless, depressed, buoyant*
Interpersonal stances	Affective stance toward another person in a specific interaction—*friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
Attitudes	Enduring, affectively colored beliefs and dispositions toward objects or persons—*liking, loving, hating, valuing, desiring*
Personality traits	Stable personality dispositions and typical behavior tendencies—*nervous, anxious, reckless, morose, hostile, jealous*

Category of Attitudes

Categorize text by:

- Type of attitude
  - Set of types (like, love, hate, etc.)
  - Commonly positive, negative, or neutral
  - Strength – number of stars
- Opinion analysis
  - The holder (source) of the attitude
  - The target (aspect) of the attitude

Why Is This Hard?

Movie reviews

- Positive—zany, rich, great, greatest
- Negative—disappointing, pathetic, worst
- These are not the only words

Many issues

- Sheer size of the language and nuances
- Negation—differences in meaning
- Sarcasm—subtle uses of language
- Ambiguity of words
- Different domains—subjects or contexts
- Mixtures of good and bad phrases

Sentiment Lexicon Approaches

Built by hand

Some employ partially automatic means

Subjectivity Cues Lexicon

LIWC—Linguistic Inquiry and Word Count

ANEW—Affective Norms for English Words

General Inquirer

Opinion Lexicon

SentiWordNet

Modeling Negation

Scope of negation

- Syntactic analysis for complex sentences
- Scope of the negation could be all words following the negation word

Negation words

- No, not, never, none, neither, nor, any word ending in “n’t”
- Other possibilities—hardly, scarcely, rarely, seldom

Intensifiers

- Very, exceedingly, less

Classification Approaches

Machine learning approach

- Document where items are labeled with appropriate sentiment attitude
- Gold standard data—appropriately labeled
- In order to get here—humans must label documents

Train a classifier

- Define features that are representative of each document
- Most frequent words and bigrams
- Must include negation modeling

Sentiment Analysis Tools

Stanford Sentiment Analyzer

- Predicts sentiment of sentences in movie reviews

SentiStrength

- Focuses on predicting positive and negative sentiments in short texts
- Lexicon based using emoji lexicons

Sentiment 140

- Focuses on predicting sentiment from tweets

Vader

- Large lexicon built from other lexicons