COM6115: Text Processing

Sentiment Analysis: Approaches

Chenghua Lin

Department of Computer Science
University of Sheffield

Learning Outcomes

By the end of the SA sessions, you will be able to:

- Explain the relevance of the topic
- Differentiate between objective and subjective texts
- List the main elements in a sentiment analysis system
- Provide a critical summary of the main approaches for the problem
- Explain how sentiment analysis systems are evaluated.

Overview

- Definition of the problem of sentiment analysis
- **Approaches to sentiment analysis**
- Evaluation of sentiment analysis approaches

Based on survey and slides by Bing Liu (University of Illinois at Chicago), 2012.

Two approaches to SA

- Lexicon-based
  - **Binary**
  - Gradable
- Corpus-based/Supervised machine learning

A simple approach to SA: lexicon-based

Use a lexicon of opinion/emotion words, like: good, bad, horrible, great, etc.

Rule-based sentiment classifier (sentence/document-level)

1. Rule-based **subjectivity classifier**: a sentence/document is subjective if it has at least n (say 2) words from the emotion words lexicon; a sentence/document is objective otherwise.
2. Rule-based **sentiment classifier**: for subjective sentences/documents, count positive and negative words/phrases in the sentence/document. If more negative than positive words/phrases, then negative; otherwise, positive (if equal, neutral).

Lexicon-based approach to SA

Rule-based sentiment classifier (feature-level)

- Assume features can be identified in previous step by information extraction techniques, e.g., battery, phone, screen.
- For each feature, count positive and negative emotion words/phrases from the lexicon.
- If more negative than positive words/phrases, then negative; otherwise, positive (if equal, neutral).

Lexicon-based approach to SA

Rule-based sentiment classifier (feature-based)

- Simple approach:
  - ◇ Input: a pair (f, s), where f is a product feature and s is a sentence that contains f.
  - ◇ Output: whether the emotion on f in s is positive, negative, or neutral.
  - ◇ **Step 1**: work on the sentence s containing f.
  - ◇ **Step 2**: select emotion words in s: $w_1, .., w_n$.
  - ◇ **Step 3**: assign orientations for these emotion words: 1 = positive, -1 = negative, 0 = neutral.
  - ◇ **Step 4**: sum up the orientation and assign the orientation to (f, s) accordingly.

- More advanced approaches split the sentence in parts, e.g., based on BUT words ("but", "except that", ...).

Lexicon-based approach to SA

Caveats

- Certain words have context-independent orientations, e.g. "great".
- Other emotion words have context-dependent orientations, e.g.
  - small power consumption = positive
  - small screen = negative
  - consume valuable resources = negative
  - consume disgusting waste = positive
- One has to deal with negation, e.g.:
  - not great = negative
  - not bad = positive
- One has to deal with intensifiers:
  - very good is more positive than good
  - extremely boring = is more negative than boring or very boring

Can store more fine-grained sentiment information in lexicon and add additional **rules**.

Two approaches to SA

- Lexicon-based
  - Binary
  - **Gradable**
- Corpus-based/Supevised machine learning

Lexicon-based approach to SA - **gradable**

Use of ranges of sentiment instead of a binary system, to deal with intensifiers like:

- absolutely, utterly, completely, totally, nearly, virtually, essentially, mainly, almost, e.g.: absolutely awful

And grading adverbs like:

- Very, little, dreadfully, extremely, fairly, hugely, immensely, intensely, rather, reasonably, slightly, unusually, e.g.: a little bit cold

Lexicon-based approach to SA - **gradable**

Rule-based gradable sentiment classifier

- Classifies general valence of a text (document-, sentence- or feature-level) based on the level of emotional content
- Level of emotional content given by:
  1. The **lexicon**: word-lists with pre-assigned emotional weights, e.g:
     Neg. dimension (C_{neg}): {-5,...,-1}, Pos. dimension (C_{pos}): {+1,...,+5}

bore	-3	careful	3
boring	-3	careless	-2
bother	-1	cares	2
brave	3	caring	3
bright	2	casual	2
brillian	2	casually	2
broke	-1	certain	2
brutal	-3	challeng	2
burden	-1	champ	2
calm	2	charit	2
care	2	charm	2
cared	2	cheat	-3
carefree	2		

Lexicon-based approach to SA - **gradable**

- Ctd:
  2. Additional **general rules** to change the original weights:

Negation rule: E.g.: "I am not good today".
Emotion(good)= +3; "not" is detected in neighbourhood (of 5 words around); so emotional valence of "good" is decreased by 1 and sign is inverted → Emotion(good) = −2

Capitalization rule: E.g. "I am GOOD today".
Emotion(good)= +3; Add +1 to positive words → Emotion(GOOD) = +4
Likewise, in "I am AWFUL today".
Emotion(awful)= −4; Add −1 to negative words → Emotion(awful) = −5

Lexicon-based approach to SA - **gradable**

Intensifier rule:

- Needs a list of intensifiers: "definitely", "very", "extremely", etc.
- Each intensifiers has a weight, e.g. Weight(very)=1; Weight(extremely)=2
- The weight is added to positive terms
- The weight is subtracted from negative terms
- E.g.: "I am feeling very good".
  Emotion(good)= +3; emotional valence of "good" increased by 1
  $\rightarrow$ Emotion(good) = +4
- E.g. "This was an extremely boring game"
  Emotion(boring)=−3; emotional valence of "boring" decreased by −2
  $\rightarrow$ Emotion(boring) = −5

Lexicon-based approach to SA - **gradable**

Diminisher rule:

- Need a list: "somewhat", "barely", "rarely", etc.
- Each intensifiers has a weight
- The weight is subtracted from positive terms
- The weight is added to negative terms
- E.g.: "I am somewhat good".
  Emotion(good)= +3; emotional valence of "good" decreased by 1
  $\rightarrow$ Emotion(good) = +2
- E.g. "This was a slightly boring game"
  Emotion(boring)=−3; emotional valence of "boring" increased by 1
  $\rightarrow$ Emotion(boring) = −2

Lexicon-based approach to SA - **gradable**

Exclamation rule: Functions like intensifiers. E.g.: "Great show!!!". Emotion(great)= +3; Weight(!!!) = 2
→ Emotion(great) = 5

Emoticon rule: Each has its own emotional weight, like an emotion word. E.g.: Emotion(☺) = +2; Emotion(☹) = −2. E.g.: "I can't believe this product ☹"
→ Emotion(☹)=−2

Lexicon-based approach to SA - **gradable**

- Final decision based on ALL emotion words:
  - If $|C_{pos}| > |C_{neg}|$ then {positive}
  - If $|C_{pos}| < |C_{neg}|$ then {negative}
  - If $|C_{pos}| = |C_{neg}|$ then {neutral}
- E.g.: "He is brilliant but boring":
  Emotion(brilliant) $= 2$; Emotion(boring) $= -3$
  $\rightarrow C_{pos} = 2$, $C_{neg} = -3$, so {negative}
- E.g.: "I am not good today":
  Emotion(good) $= -2$
  $\rightarrow C_{pos} = 0$, $C_{neg} = -2$, so {negative}
- E.g.: "I am not GOOD today": (Emotion(good)=3) $\rightarrow$???
- E.g.: "I am so surprised by this product!!! ☹": (Emotion(☹)=-2) $\rightarrow$???

Lexicon-based approach to SA

Advantages:

- Works effectively with different texts: forums, blogs, etc.
- Language independent - as long as an up-to-date lexicon of emotion words is available
- Doesn't require data for training
- Can be extended with additional lexica, e.g. for new emotion words/symbols as they become popular, esp. in social media

Disadvantages:

- Requires a lexicon of emotion words, which should be fairly comprehensive, covering new words, abbreviations (LOL, m8, etc.), misspelled words, etc.

E.g.: In a dataset from MySpace, 95% of comments contained at least one spelling error!

Lexica of emotion words/phrases

For both binary and gradable approaches, how to obtain lexica of emotion words?

Task: Collect relevant words/phrases that can be used to express sentiment. Determine the emotion of these subjective word/phrases.

- **Manually**: word lists with pre-assigned emotional weights
- Semi-automatically
  - **Dictionary-based**: find synonyms/antonyms of seed emotion words in dictionaries like WordNet
  - **Corpus-based**: find synonyms/antonyms of seed emotion words in corpora

Lexica of emotion words/phrases (ctd)

Mostly adjectives

- Positive: e.g.: honest, important, mature, large, patient, ...
- Negative: harmful, hypocritical, inefficient, insecure

Verbs

- Positive: praise, love
- Negative: blame, criticize

Nouns

- Positive: pleasure, enjoyment
- Negative: pain, criticism

Phrases (esp. for collocations, but also alternative to having intensifiers weighted separately)

- Positive: high intelligence, low cost
- Negative: little variation, many problems

Lexica of emotion words/phrases (ctd)

Semi-automatically created resources, such as:

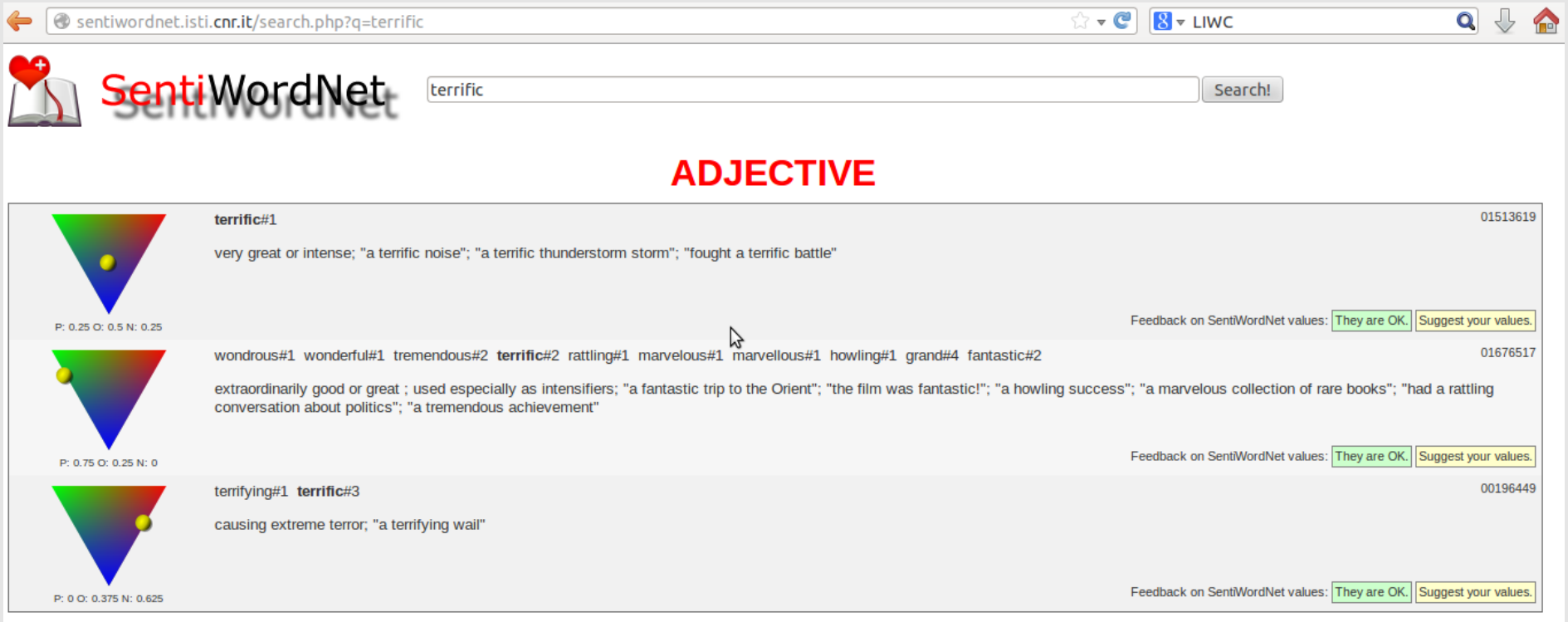
- **SentiWordNet**: Wordnet is a database with words grouped into sets of synonyms (synsets), and organised by semantic relations between them: synonyms, antonyms, hypernyms, etc. SentiWordNet is a version of it with one of three sentiment scores for each synset: positivity, negativity, objectivity.

Manually created resources, such as:

- **Linguistic Inquiry and Word Count (LIWC) lexicon**: made by psychologists with lists of words with various emotional and other dimensions.
- **General Inquirer**: terms with various types of positive or negative semantic orientation.

Lexica of emotion words/phrases (ctd)

SentiWordNet



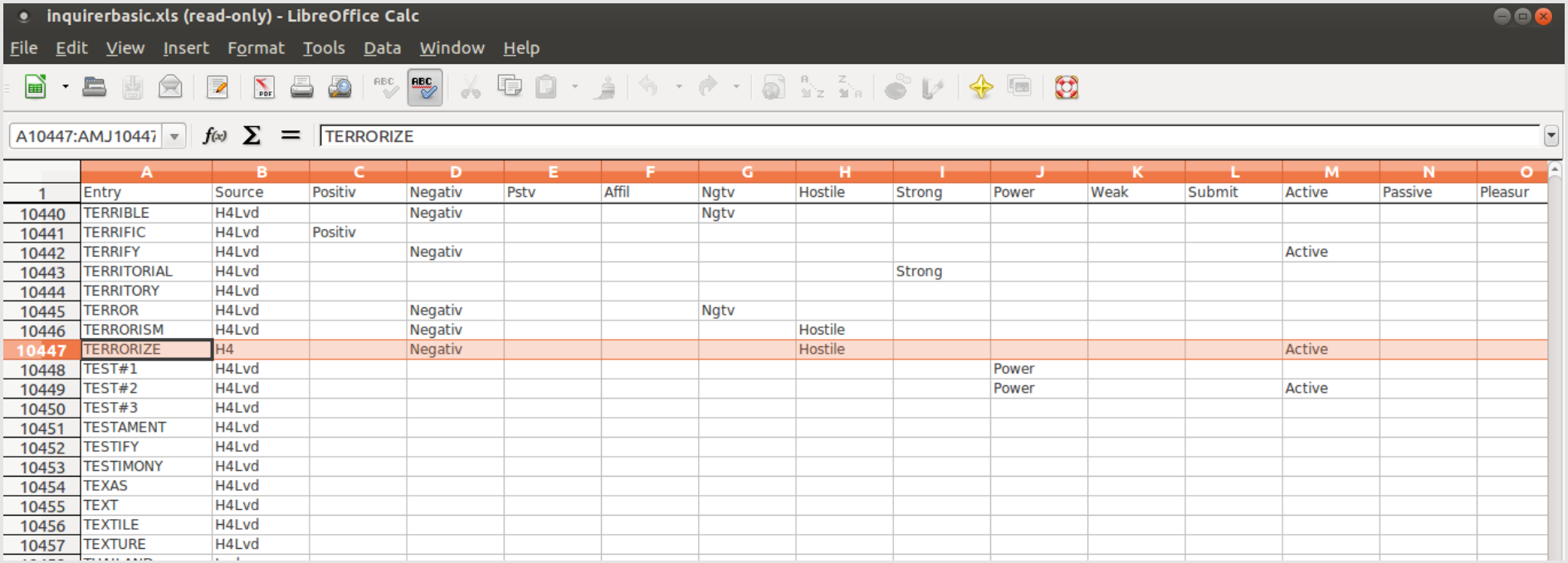
Lexica of emotion words/phrases (ctd)

Linguistic Inquiry and Word Count lexicon

Category	Abbrev	Examples	Words In Category
Psychological Processes			
Social processes	social	Mate, talk, they, child	455
Family	family	Daughter, husband, aunt	64
Friends	friend	Buddy, friend, neighbor	37
Humans	human	Adult, baby, boy	61
Affective processes	affect	Happy, cried, abandon	915
Positive emotion	posemo	Love, nice, sweet	406
Negative emotion	negemo	Hurt, ugly, nasty	499
Anxiety	anx	Worried, fearful, nervous	91
Anger	anger	Hate, kill, annoyed	184
Sadness	sad	Crying, grief, sad	101
Cognitive processes	cogmech	cause, know, ought	730
Insight	insight	think, know, consider	195

Lexica of emotion words/phrases (ctd)

General Inquirer: words classified in many categories, including: positive (1,915) and negative (2,291).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Entry	Source	Positiv	Negativ	Pstv	Affil	Ngtv	Hostile	Strong	Power	Weak	Submit	Active	Passive	Pleasur
10440	TERRIBLE	H4Lvd		Negativ			Ngtv								
10441	TERRIFIC	H4Lvd	Positiv												
10442	TERRIFY	H4Lvd		Negativ									Active		
10443	TERRITORIAL	H4Lvd							Strong						
10444	TERRITORY	H4Lvd													
10445	TERROR	H4Lvd		Negativ			Ngtv								
10446	TERRORISM	H4Lvd		Negativ				Hostile							
10447	TERRORIZE	H4		Negativ				Hostile					Active		
10448	TEST#1	H4Lvd								Power					
10449	TEST#2	H4Lvd								Power			Active		
10450	TEST#3	H4Lvd													
10451	TESTAMENT	H4Lvd													
10452	TESTIFY	H4Lvd													
10453	TESTIMONY	H4Lvd													
10454	TEXAS	H4Lvd													
10455	TEXT	H4Lvd													
10456	TEXTILE	H4Lvd													
10457	TEXTURE	H4Lvd													

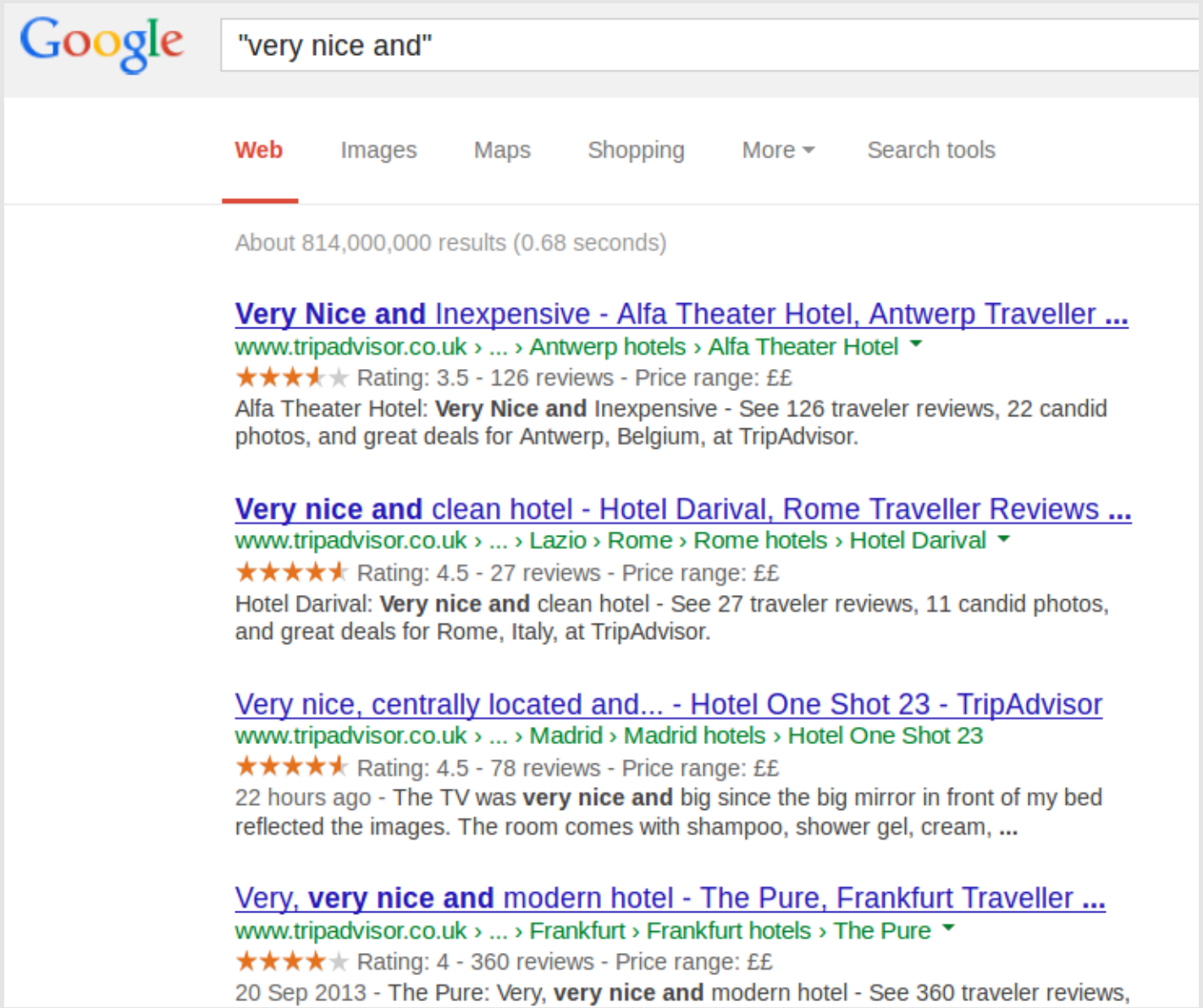
Free dictionary:
`http://www.wjh.harvard.edu/~inquirer/homecat.htm`

Lexica of emotion words/phrases (ctd)

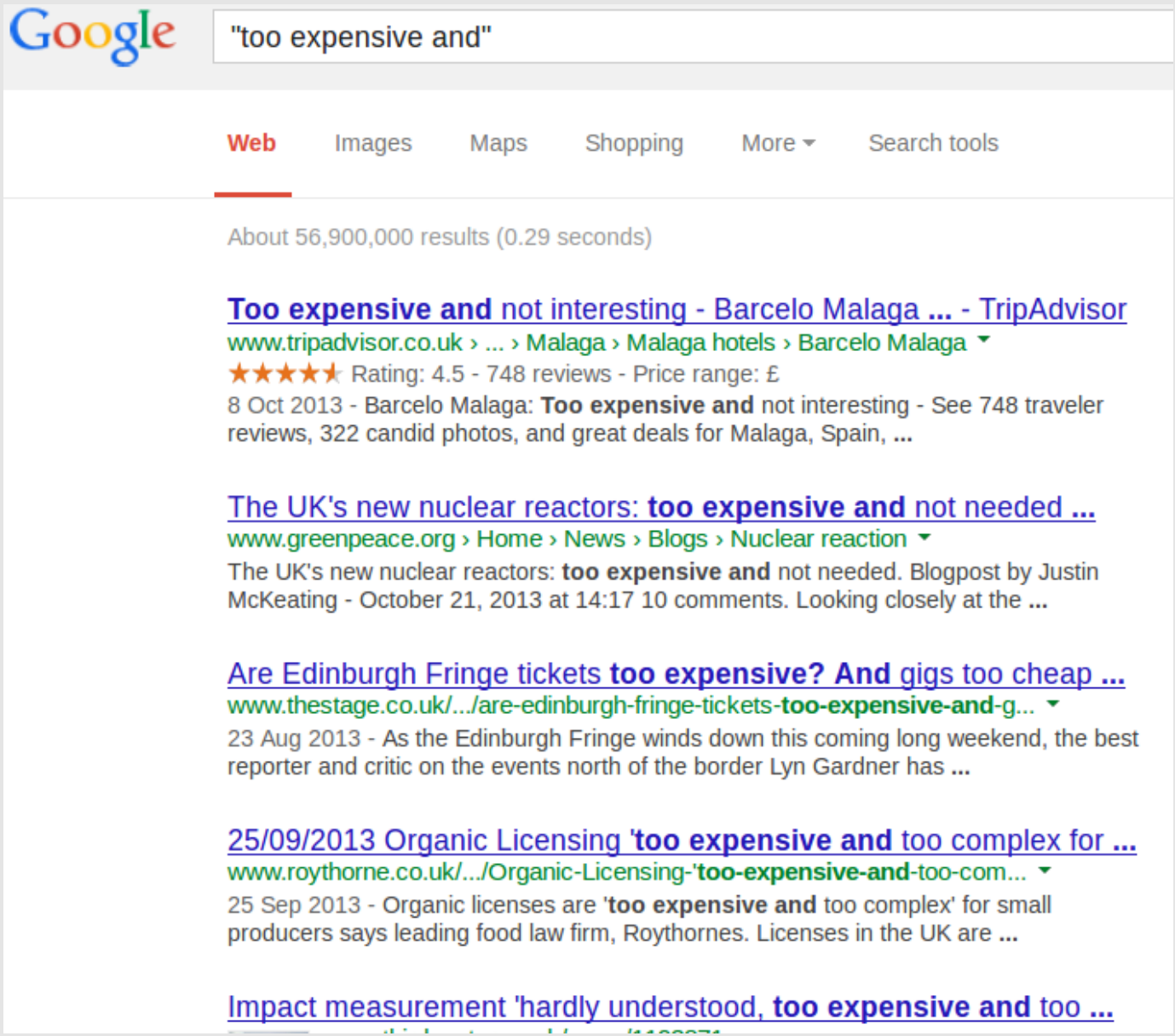
Semi-automatically created from seed words: start with **seed positive and negative words**:

- Search for synonyms/antonyms in dictionaries like WordNet; OR
- Build **patterns** from seed words/phrases to search on large corpora, like the Web:
  - "beautiful and" (+)
  - "low cost but" (-)
  - "very nice and " (+)

Lexica of emotion words/phrases (ctd) - from corpora



Lexica of emotion words/phrases (ctd) - from corpora



Two approaches to SA

- Lexicon-based
  - Binary
  - Gradable
- **Corpus-based/Supervised machine learning**

A corpus-based approach to SA

Idea: Mostly "supervised learning": **corpora** of examples **annotated with sentiment** are used with machine learning algorithms to learn a classifier for each sentence/document. Corpora can be built:

- Manually: reliable, can be used as gold-standards
- From crowd-annotated resources, like Amazon Product Reviews (1-5 stars); Rotten Tomatoes, complaints.com, bitterlemons.com

Corpus: a collection of text segments (e.g. webpages, blog posts, reviews, tweets, etc) with humanly-annotated emotional indicators (e.g. positive, negative, etc).

E.g.: "If you are reading this because it is your darling fragrance, please wear it at home exclusively and tape the windows shut." $\rightarrow$ {negative}

A corpus-based approach to SA - Corpora

Examples of corpora:

- Subjectivity corpus
  - 10,000 sentences: subjective/objective
  - Objective: IMDB plot summaries
  - Subjective: Rotten Tomatoes website.
- "Movie Review" corpus (Pang, Lee and Vaithyanathan, 2002):
  - 2,000 movie reviews (equal number of positive/negative)
  - Source: IMDB
- Many more:
  `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`

A corpus-based approach to SA - Features

Mostly words, but also other linguistic traits describing positive/negative examples:

- Words (unigrams)
- n-grams (sequences of n words)
- Emotions from words/phrases extracted from dictionaries
- Part-of-speech (POS) tags
- Syntactic patterns (e.g. sequences of POS tags)
- Language model scores: similarities to positive/negative corpora
- Negations

All automatically extracted from the corpus.

A corpus-based approach to SA - Machine Learning

Two steps:

1. Subjectivity classifier: first run binary classifier to identify and then eliminate objective segments
2. Sentiment classifier with remaining segments: learn how to combine and weight different attributes to make predictions. E.g. Naive Bayes

Pre-processing of corpus similar to IR:

- Remove HTML or other tags
- Remove stopwords
- Perform word stemming/lemmatisation
- etc.

Extra reading

Bing Liu and Lei Zhang (2012). A survey on opinion mining and sentiment analysis. Kluwer Academic Publishers:
`http://www.cs.uic.edu/~lzhang3/paper/opinion_survey.pdf`