

ELEG 6913: Machine Learning for Big Data

Fall 2016

Lecture 12: Sentiment Analysis and Named Entity Recognition

Dr. Xishuang Dong

Outline

- **Natural Language Processing**
- **Sentiment Analysis**
- **Named Entity Recognition**
- **Summary**

(Acknowledgment: some parts of the slides are from Jenny Rose Finkel, Zornitsa Kozareva, Dan Jurafsky, and various other sources. The copyright of those parts belongs to their original owners.)

Outline

- **Natural Language Processing**
- Sentiment Analysis
- Named Entity Recognition
- Summary

Natural Language Processing (NLP) Problems

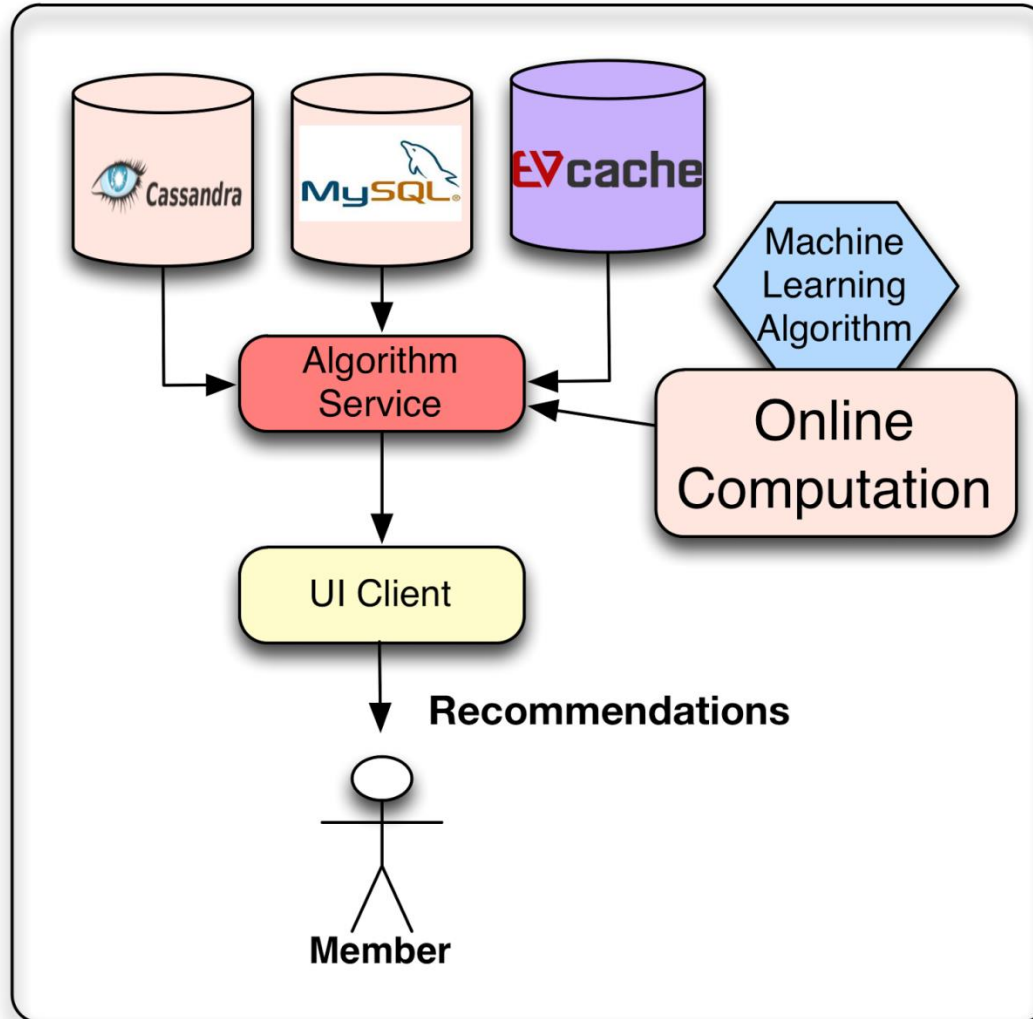
- **Text Classification**
- **Text Clustering**
- **Sentiment Analysis**
- **Named Entity Recognition (NER)**
- **... ..**

Natural Language Processing (NLP) Problems

- Text Classification
- Text Clustering
- Sentiment Analysis
- Named Entity Recognition (NER)
-

NLP Applications

- **Recommendation System**





Outline

- Natural Language Processing
- **Sentiment Analysis**
- Named Entity Recognition
- Summary

Positive or negative movie review?

 • **Unbelievably disappointing**

 • **Full of zany characters and richly applied satire, and some great plot twists**

 • **This is the greatest screwball comedy ever filmed**

 • **It was pathetic. The worst part about it was the boxing scenes.**

Google Product Search



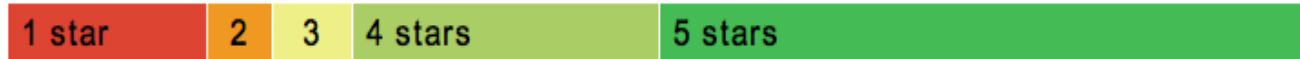
HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner

\$89 online, \$100 nearby ★★★★★ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 shi

Reviews

Summary - Based on 377 reviews




What people are saying

ease of use		"This was very easy to setup to four computers."
value		"Appreciate good quality at a fair price."
setup		"Overall pretty easy setup."
customer service		"I DO like honest tech support people."
size		"Pretty Paper weight."
mode		"Photos were fair on the high quality mode."
colors		"Full color prints came out with great quality."

Bing Shopping

HP Officejet 6500A E710N Multifunction Printer

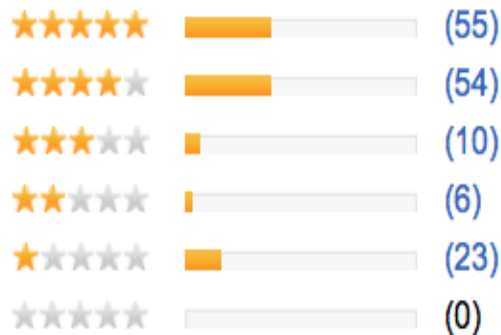
[Product summary](#) [Find best price](#) **Customer reviews** [Specifications](#) [Related items](#)



\$121.53 - \$242.39 (14 stores)

☐ Compare

Average rating ★★★★★ (144)



Most mentioned



Show reviews by source

Best Buy (140)
CNET (5)
Amazon.com (3)

Target Sentiment on Twitter

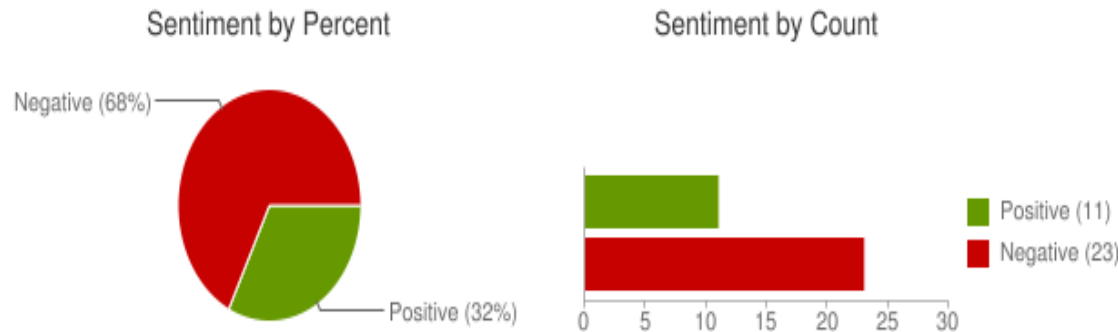
Type in a word and we'll highlight the good and the bad

"united airlines"

Search

[Save this search](#)

Sentiment analysis for "united airlines"



iljacobson: OMG... Could @United airlines have worse customer service? W8g now 15 minutes on hold 4 questions about a flight 2DAY that need a human.
Posted 2 hours ago

12345clumsy6789: I hate United Airlines Ceiling!!! Fukn impossible to get my conduit in this damn mess! ?
Posted 2 hours ago

EMLandPRGbelgiu: EML/PRG fly with Q8 united airlines and 24seven to an exotic destination. <http://t.co/Z9QloAjF>
Posted 2 hours ago

CountAdam: FANTASTIC customer service from United Airlines at XNA today. Is tweet more, but cell phones off now!
Posted 4 hours ago

Sentiment analysis has many other names

- **Opinion extraction**
- **Opinion mining**
- **Sentiment mining**
- **Subjectivity analysis**

Why sentiment analysis?

- ***Movie:*** is this review positive or negative?
- ***Products:*** what do people think about the new iPhone?
- ***Public sentiment:*** how is consumer confidence? Is despair increasing?
- ***Politics:*** what do people think about this candidate or issue?
- ***Prediction:*** predict election outcomes or market trends from sentiment

Scherer Typology of Affective States

- **Emotion:** brief organically synchronized ... evaluation of a major event
 - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood:** diffuse non-caused low-intensity long-duration change in subjective feeling
 - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances:** affective stance toward another person in a specific interaction
 - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*

Scherer Typology of Affective States

- **Attitudes:** enduring, affectively colored beliefs, dispositions towards objects or persons
 - *liking, loving, hating, valuing, desiring*
- **Personality traits:** stable personality dispositions and typical behavior tendencies
 - *nervous, anxious, reckless, morose, hostile, jealous*

Sentiment Analysis

- **Sentiment analysis is the detection of attitudes “enduring, affectively colored beliefs, dispositions towards objects or persons”**
 - 1. Holder (source) of attitude**
 - 2. Target (aspect) of attitude**
 - 3. Type of attitude**
 - From a set of types
 - *Like, love, hate, value, desire, etc.*
 - Or (more commonly) simple weighted **polarity**:
 - *positive, negative, neutral, together with strength*
 - 4. Text containing the attitude**
 - Sentence or entire document

Sentiment Analysis

- **Simplest task:**
 - Is the attitude of this text positive or negative?
- **More complex:**
 - Rank the attitude of this text from 1 to 5
- **Advanced:**
 - Detect the target, source, or complex attitude types

Sentiment Classification in Movie Reviews

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL, 271-278

- **Sentiment Polarity Detection:**
 - Is an IMDB movie review positive or negative?
- **Data: *Polarity Data 2.0*:**
 - <http://www.cs.cornell.edu/people/pabo/movie-review-data>

IMDB Data



when _star wars_ came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image . [...]

when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point .

cool .

october sky offers a much simpler image— that of a single white dot , traveling horizontally across the night sky . [. . .]



“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing . it’s not just because this is a brian depalma film , and since he’s a great director and one who’s films are always greeted with at least some fanfare .

and it’s not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents .

Baseline Algorithm

- **Feature Extraction**
- **Classification using different classifiers**
 - **Naive Bayes**
 - **Maximum Entropy**
 - **SVM**
 - **Logistic Regression**
 - **... ..**

Extracting Features for Sentiment Classification

- **How to handle negation**
 - I **didn't** **like** this movie
- VS
- I **really** **like** this movie
- **Which words to use?**
 - **Only adjectives**
 - **All words**
 - **All words turns out to work better.**

Negation

- Add **NOT_** to every word between negation and following punctuation:

didn' t like this movie , but I

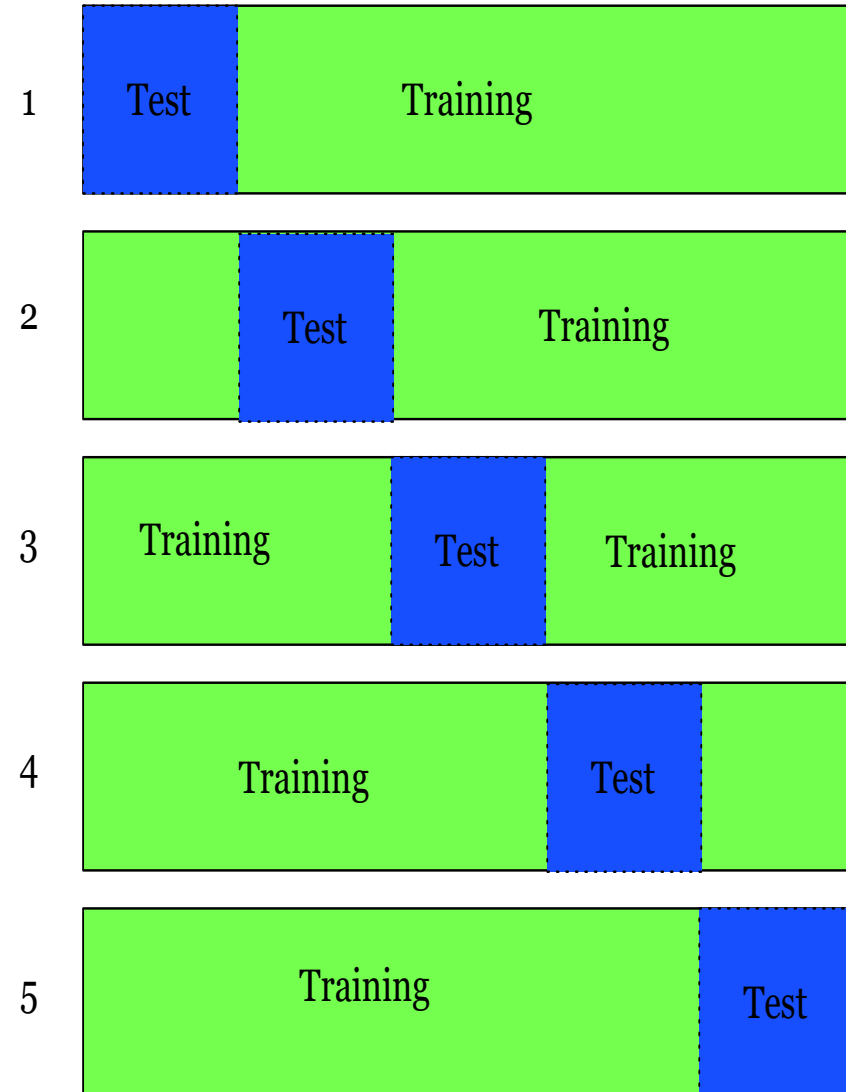


didn' t NOT_like NOT_this NOT_movie but I

Cross-Validation

- Break up data into 10 folds
 - (Equal positive and negative inside each fold?)
- For each fold
 - Choose the fold as a temporary test set
 - Train on 9 folds, compute performance on the test fold
- Report average performance of the 10 runs

Iteration



What makes reviews hard to classify?

- **Subtlety:**
 - **Perfume review in *Perfumes*:**
 - **“If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”**
 - **Dorothy Parker on Katherine Hepburn**
 - **“She runs the gamut of emotions from A to B”**

Thwarted Expectations and Ordering Effects

- “This film should be **brilliant**. It sounds like a **great** plot, the actors are **first grade**, and the supporting cast is **good** as well, and Stallone is attempting to deliver a good performance. However, it **can’t hold up**.”
- Well as usual Keanu Reeves is nothing special, but surprisingly, the **very talented** Laurence Fishbourne is **not so good** either, I was surprised.

The General Inquirer

- **Home page:** <http://www.wjh.harvard.edu/~inquirer>
- **List of Categories:**
<http://www.wjh.harvard.edu/~inquirer/homecat.htm>
- **Spreadsheet:**
<http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls>
- **Categories:**
 - **Positiv (1915 words) and Negativ (2291 words)**
 - **Strong vs Weak, Active vs Passive, Overstated vs Understated**
 - **Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc**
- **Free for Research Use**

SentiWordNet

- Home page: <http://sentiwordnet.isti.cnr.it/>
- All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness
- [estimable(J,3)] “may be computed or estimated”

Pos 0 Neg 0 Obj 1

- [estimable(J,1)] “deserving of respect or high regard”

Pos .75 Neg 0 Obj .25

Other sentiment feature:

Logical negation

- Is logical negation (*no*, *not*) associated with negative sentiment?
- Potts experiment:
 - Count negation (*not*, *n't*, *no*, *never*) in online reviews
 - Regress against the review rating

Intuition for identifying word polarity

- Adjectives conjoined by “*and*” have same polarity
 - Fair and legitimate, corrupt and brutal
 - *fair and brutal, *corrupt and legitimate
- Adjectives conjoined by “*but*” do not
 - fair but brutal

Feature Extractions: step 1

- Label seed set of 1336 adjectives
- 657 positive
 - adequate, central, clever, famous, intelligent, remarkable, reputed, sensitive, slender, thriving...
- 679 negative
 - contagious, drunken, ignorant, lanky, listless, primitive, strident, troublesome, unresolved, unsuspecting...

Feature Extractions: step 2

- Expand seed set to conjoined adjectives



"was nice and"

[Nice location in Porto and the front desk staff was **nice and helpful** ...](#)

[www.tripadvisor.com/ShowUserReviews-g189180-d206904-r12068...](#) 

Mercure Porto Centro: Nice location in Porto and the front desk staff **was nice and helpful** - See traveler reviews, 77 candid photos, and great deals for Porto, ...

nice, helpful

[If a girl was **nice and classy** but had some vibrant purple dye in ...](#)

[answers.yahoo.com › Home › All Categories › Beauty & Style › Hair](#) 

4 answers - Sep 21

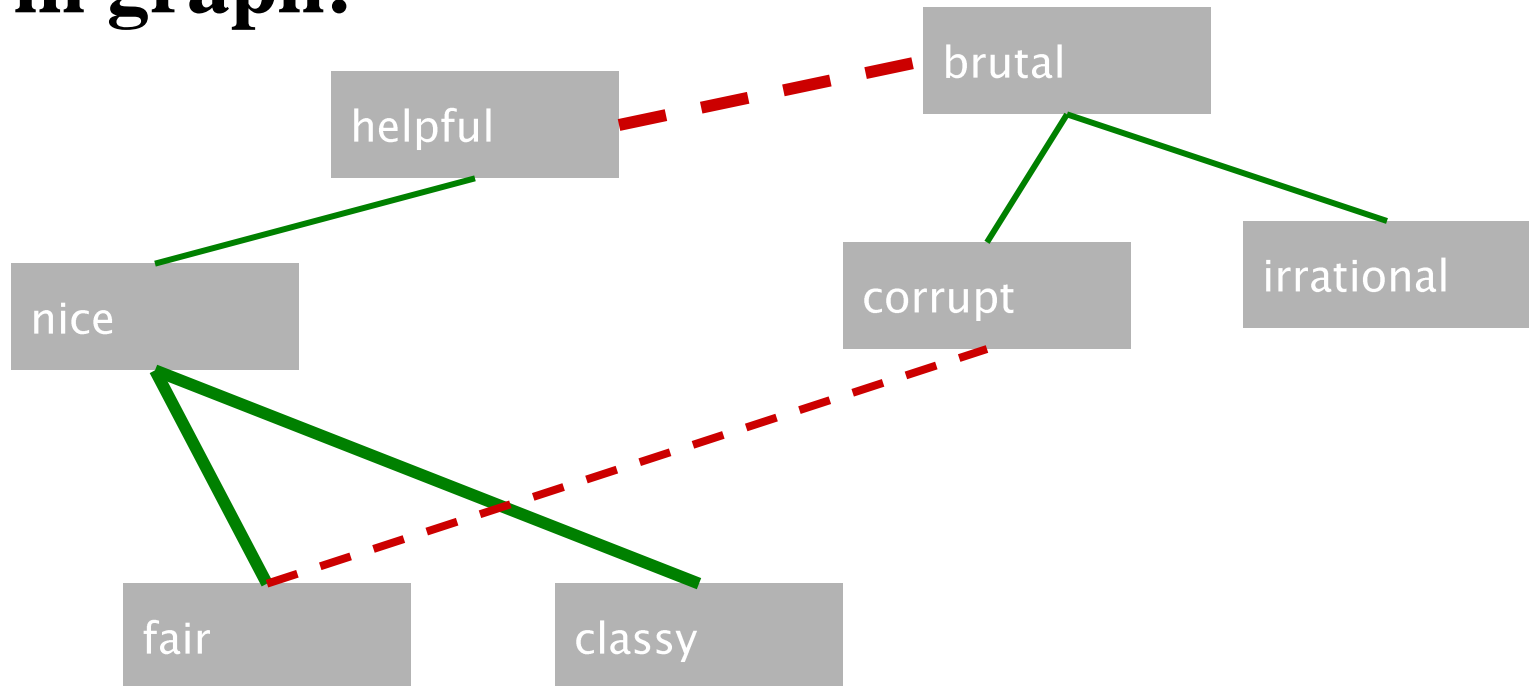
Question: Your personal opinion or what you think other people's opinions might ...

Top answer: I think she would be cool and confident like katy perry :)

nice, classy

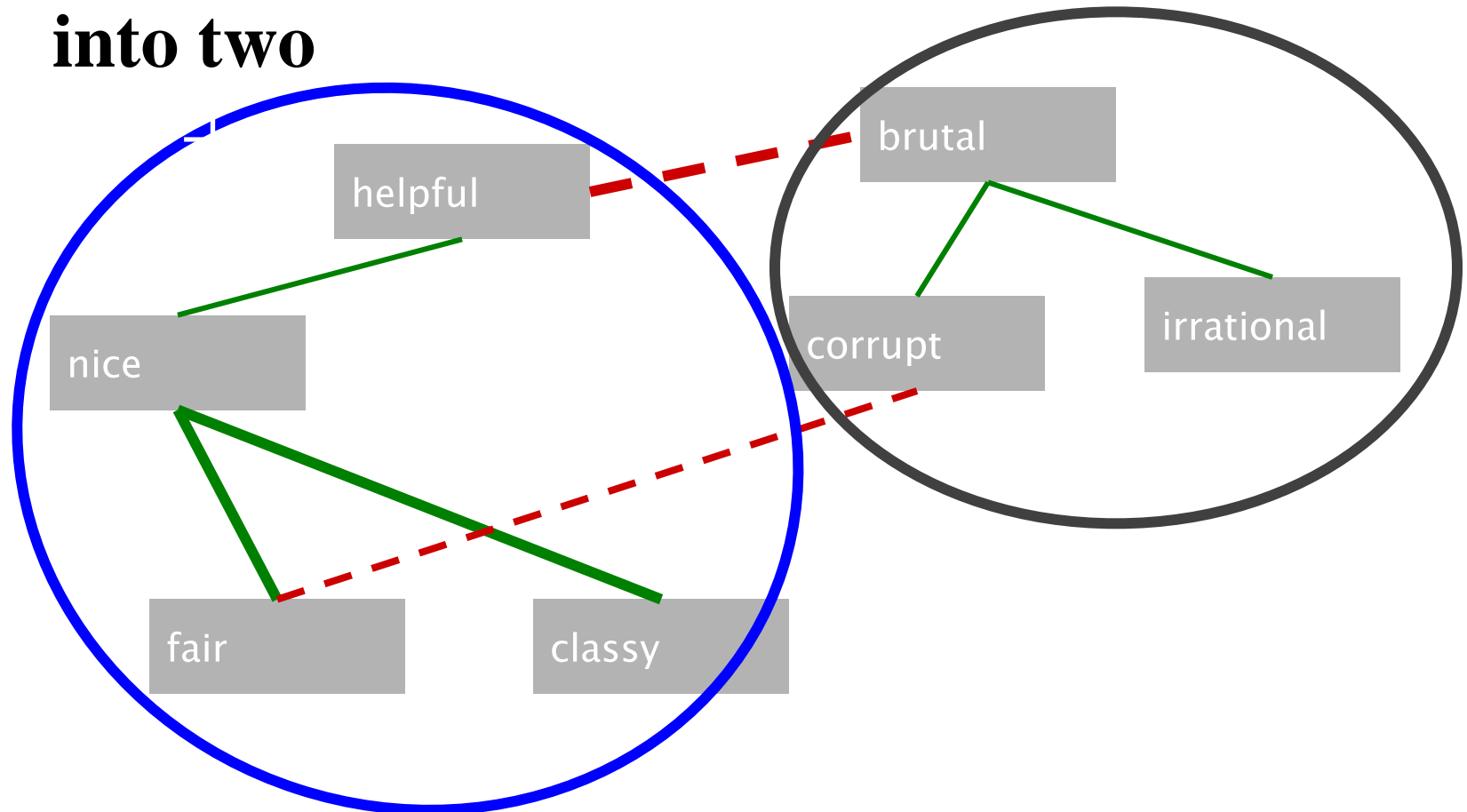
Feature Extractions: step 3

- Supervised classifier assigns “polarity similarity” to each word pair, resulting in graph:



Feature Extractions: step 4

- **Clustering for partitioning the graph into two**



Using WordNet to learn polarity

- **WordNet: online thesaurus (covered in later lecture).**
- **Create positive (“good”) and negative seed-words (“terrible”)**
- **Find Synonyms and Antonyms**
 - **Positive Set: Add synonyms of positive words (“well”) and antonyms of negative words**
 - **Negative Set: Add synonyms of negative words (“awful”) and antonyms of positive words (“evil”)**
- **Repeat, following chains of synonyms**

Summary on Learning Lexicons

- **Advantages:**
 - **Can be domain-specific**
 - **Can be more robust (more words)**
- **Intuition**
 - **Start with a seed set of words ('good', 'poor')**
 - **Find other words that have similar polarity:**
 - **Using “and” and “but”**
 - **Using words that occur nearby in the same document**
 - **Using WordNet synonyms and antonyms**
 - **Use seeds and semi-supervised learning to induce lexicons**

How to measure polarity of a phrase?

- Positive phrases co-occur more with *“excellent”*
- Negative phrases co-occur more with *“poor”*
- But how to measure co-occurrence?

Pointwise Mutual Information

- **Mutual information** between 2 random variables X and Y

$$I(X, Y) = \sum_x \sum_y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- **Pointwise mutual information:**
 - How much more do events x and y co-occur than if they were independent?

$$\text{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

Pointwise Mutual Information

- **Pointwise mutual information:**
 - How much more do events x and y co-occur than if they were independent?

$$\text{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- **PMI between two words:**
 - How much more do two words co-occur than if they were independent?

$$\text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1, \text{word}_2)}{P(\text{word}_1)P(\text{word}_2)}$$

Finding sentiment of a sentence

- Important for finding aspects or attributes
 - Target of sentiment

“The **food** was **great** but the **service** was awful”

Finding aspect/attribute/target of sentiment

- The aspect name may not be in the sentence
- For restaurants/hotels, aspects are well-understood
- Supervised classification
 - Hand-label a small corpus of restaurant review sentences with aspect
 - food, d écor, service, value, NONE
 - Train a classifier to assign an aspect to a sentence
 - “Given this sentence, is the aspect *food*, *d écor*, *service*, *value*, or *NONE*”

Outline

- Natural Language Processing
- Sentiment Analysis
- **Named Entity Recognition**
- Summary

Named Entity Recognition and Classification

<PER>**Prof. Jerry Hobbs**</PER> taught CS544 during
<DATE>**February 2010**</DATE>.
<PER>**Jerry Hobbs**</PER> killed his daughter in
<LOC>**Ohio**</LOC>.
<ORG>**Hobbs corporation**</ORG> bought
<ORG>**FbK**</ORG>.

Named Entity Recognition and Classification

- Identify mentions in text and classify them into a predefined set of categories of interest:
 - Person Names: **Prof. Jerry Hobbs, Jerry Hobbs**
 - Organizations: **Hobbs corporation, FbK**
 - Locations: **Ohio**
 - Date and time expressions: **February 2010**
 - E-mail: **mkg@gmail.com**
 - Web address: **www.usc.edu**
 - Names of drugs: **paracetamol**
 - Names of ships: **Queen Marry**
 - Bibliographic references:
 - ...

Knowledge NER vs. Learning NER

Knowledge Engineering

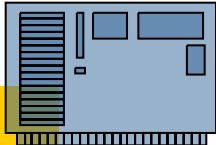


- + very precise (hand-coded rules)
- + small amount of training data
- expensive development & test cycle
- domain dependent
- changes over time are hard

Learning Systems



- + higher recall
- + no need to develop grammars
- + developers do not need to be experts
- + annotations are cheap
- require lots of training data



Rule Based NER

- **Create regular expressions to extract:**
 - **Telephone number**
 - **(***)-***-******
 - **E-mail**
 - *******@******
 - **Capitalized names**
 - **No special chars like “*, \$, @, #,”**

Rule Based NER

- **Regular expressions provide a flexible way to match strings of text, such as particular characters, words, or patterns of characters**

Suppose you are looking for a word that:

- 1. starts with a capital letter “P”**
- 2. is the first word on a line**
- 3. the second letter is a lower case letter**
- 4. is exactly three letters long**
- 5. the third letter is a vowel**

Rule Based NER

- **Regular expressions provide a flexible way to match strings of text, such as particular characters, words, or patterns of characters**

the regular expression would be “`^P[a-z][aeiou]`” where

`^` - indicates the beginning of the string

`[a-z]` – any letter in range a to z

`[aeiou]` – any vowel

Perl RegEx

- `\w` (word char) any alpha-numeric
- `\d` (digit char) any digit
- `\s` (space char) any whitespace
- `.` (wildcard) anything
- `\b` word bounday
- `^` beginning of string
- `$` end of string
- `?` For 0 or 1 occurrences
- `+` for 1 or more occurrences
- specific range of number of occurrences: `{min,max}`.
 - `A{1,5}` One to five A's.
 - `A{5,}` Five or more A's
 - `A{5}` Exactly five A's

Rule Based NER

- **Create regular expressions to extract:**

- Telephone number
- E-mail
- Capitalized names

blocks of digits separated by hyphens

RegEx = (\d+\-)+\d+

- matches valid phone numbers like 900-865-1125 and 725-1234
- incorrectly extracts social security numbers 123-45-6789
- fails to identify numbers like 800.865.1125 and (800)865-CARE

Improved RegEx = (\d{3}[-.\ ()])\{1,2\}[\dA-Z]\{4\}

Rule Based NER

- **Create rules to extract locations**
 - Capitalized word + {city, center, river} indicates location

Ex. *New York city*

Hudson river

- Capitalized word + {street, boulevard, avenue} indicates location

Ex. *Fifth avenue*

Why simple things would not work?

- Capitalization is a strong indicator for capturing proper names, but it can be tricky:
 - first word of a sentence is capitalized
 - sometimes titles in web pages are all capitalized
 - nested named entities contain non-capital words
“university of southern california” is
Organization
 - all nouns in German are capitalized

Why simple things would not work?

- The same entity can have multiple variants of the same proper name

Zornitsa Kozareva

prof. Kozareva

Zori



- Proper names are ambiguous

Jordan the *person* vs. Jordan the *location*

JFK the *person* vs. JFK the *airport*

May the *person* vs. May the *month*

Learning System

- *Supervised* learning
 - labeled training examples
 - methods: k-Nearest Neighbors, SVM, ...
 - example: NE recognition, POS tagging, Parsing
- *Unsupervised* learning
 - labels must be automatically discovered
 - method: clustering
 - example: NE disambiguation, text classification

Learning System

- *Semi-supervised* learning
 - small percentage of training examples are labeled, the rest is unlabeled
 - methods: bootstrapping, active learning, co-training, self-training
 - example: NE recognition, POS tagging, Parsing, ...

Machine Learning Based NER

Adam_B Smith_I works_O for_O IBM_B ,_O London_B ._O

- **NED:** Identify named entities using BIO tags
 - B: beginning of an entity
 - I: continues the entity
 - O: word outside the entity

Machine Learning Based NER

Adam_B-PER Smith_I-PER works_O for_O IBM_B-ORG ,_O London_B-LOC ._O

- **NED**: Identify named entities using BIO tags
 - B beginning of an entity
 - I continues the entity
 - O word outside the entity
- **NEC**: Classify into a predefined set of categories
 - Person names
 - Organizations (companies, governmental organizations, etc.)
 - Locations (cities, countries, etc.)
 - Miscellaneous (movie titles, sport events, etc.)

NER Data/Bake-Offs

- **CoNLL-2002 and CoNLL-2003 (British newswire)**
 - **Multiple languages: Spanish, Dutch, English, German**
 - **4 entities: Person, Location, Organization, Misc**
- **MUC-6 and MUC-7 (American newswire)**
 - **7 entities: Person, Location, Organization, Time, Date, Percent, Money**
- **ACE**
 - **5 entities: Location, Organization, Person, FAC, GPE**
- **BBN (Penn Treebank)**
 - **22 entities: Animal, Cardinal, Date, Disease, ...**

Features for NE Detection (1)

Adam
Smith
Works
for
IBM
in
London

Features for NE Detection (1)

Adam
Smith
Works
for
IBM
in
London
.

Adam
Smith

.....

fp

Features for NE Detection (1)

Adam
Smith
Works
for
IBM
in
London
.

Adam, null, null, null, Smith, works, for
Smith, Adam, null, null, works, for, IBM

.....

fp, London, in, IBM, null, null, null

Features for NE Detection (2)

Adam
for,1,0

Smith
IBM,1,0

Works

for

IBM

in

London

.

Adam, null, null, null, Smith, works,

Smith, Adam, null, null, works, for,

.....

fp, London, in, IBM, null, null, null,0,0

Features for NE Detection (3)

- **Orthographic (binary and not mutually exclusive)**

initial-caps

roman-number

acronym

single-char

all-caps

contains-dots

lonely-initial

*functional-word**

all-digits

contains-hyphen

punctuation-mark

URL

- **Word-Type Patterns:**

functional

capitalized

lowercased

punctuation mark

quote

other

- **Left Predictions**

- the tag predicted in the current classification for W-3, W-2, W-1

- **Part-of-speech (POS) tag** (when available)

The more useful features you incorporate, the more powerful your learner gets

Features for NE Classification (1)

- **Contextual**
 - current word W_0
 - words around W_0 in $[-3, \dots, +3]$ window
- **Part-of-speech tag** (when available)
- **Bag-of-Words**
 - words in $[-5, \dots, +5]$ window
- **Trigger words**
 - for person (*Mr, Miss, Dr, PhD*)
 - for location (*city, street*)
 - for organization (*Ltd., Co.*)
- **Gazetteers**
 - geographical
 - first name
 - surname
 - company names

Features for NE Classification (2)

- **Length in words of the entity being classified**
- **Pattern of the entity with regard to the type of constituent words**
- **For each class**
 - **whole NE is in gazetteer**
 - **any component of the NE appears in gazetteer**
- **Suffixes (length 1 to 4)**
 - **each component of the NE**
 - **whole NE**

Outline

- Natural Language Processing
- Sentiment Analysis
- Named Entity Recognition
- **Summary**

Summary on Sentiment Analysis

- **Generally modeled as classification or regression task**
 - **predict a binary or ordinal label**
- **Features:**
 - **Negation is important**
 - **Using all words (in naïve bayes) works well for some tasks**
 - **Finding subsets of words may help in other tasks**
 - **Hand-built polarity lexicons**
 - **Use seeds and semi-supervised learning to induce lexicons**

Summary on NER

- **Named-entity recognition**
 - **entity identification, entity chunking and entity extraction**
- **A subtask of information extraction**
- **Seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.**
- **Knowledge Base**

Deep Learning for Big Data

Thank you!

Q&A