Text Summarization - LexRank

Pawan Goyal

CSE, IIT Kharagpur

Week 11, Lecture 1

Text Summarization

What is a summary?

A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). (*Hovy, 2008*)

Text Summarization

What is a summary?

A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). (*Hovy, 2008*)

What is text summarization?

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user or task. (*Mani and MayBury, 2001*)

Text Summarization

What is a summary?

A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). (*Hovy, 2008*)

What is text summarization?

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user or task. (*Mani and MayBury, 2001*)

Humans have an incredible capacity to condense information down to the critical bit.

> *"He said he is against it."*

Calvin Coolidge, on being asked what a clergyman preaching on sin said.

Goal of a Text Summarization System

To give an overview of the original document in a shorter period of time.

Automatic Text Summarization

Goal of a Text Summarization System

To give an overview of the original document in a shorter period of time.

Summarization Applications

- *outlines or abstracts* of any document, news article etc.
- *summaries* of email threads
- *action items* from a meeting
- *simplifying* text by compressing sentences

Application: Generating Snippets

Robert O'Neill taking credit for killing Osama bin Laden sparks debate
Hindustan Times - 1 hour ago
Some special operations service members and veterans are unhappy that one of their own has taken credit publicly for killing Osama bin Laden.

It's been special knock as wait has been long: Rayudu
Hindustan Times - 1 hour ago
An elated Ambati Rayudu said Friday that his maiden hundred in international cricket will certainly be a "special one" as it took a long time to come.

Application: Generating Snippets

Robert O'Neill taking credit for killing Osama bin Laden sparks debate
Hindustan Times - 1 hour ago
Some special operations service members and veterans are unhappy that one of their own has taken credit publicly for killing Osama bin Laden.

It's been special knock as wait has been long: Rayudu
Hindustan Times - 1 hour ago
An elated Ambati Rayudu said Friday that his maiden hundred in international cricket will certainly be a "special one" as it took a long time to come.

what is the relation between pressure and velocity

Web Images Videos News More Search tools

About 1,10,00,000 results (0.49 seconds)

fluid dynamics - Relation between pressure, velocity and ...
physics.stackexchange.com/.../**relation-between-pressure-velocity**-and-ar...
In a nozzle, the exit **velocity** increases as per continuity equation as given by Bernoulli equation (incompressible fluid). **Pressure** is inversely proportional to ...

Chapter 9: Fluid Dynamics
francesa.phy.cmich.edu/people/andy/physics110/book/.../Chapter9.htm
From practical experience we know that the velocity of fluid through the small ... we found a qualitative **relationship between pressure and velocity** in a fluid flow.

Bernoulli's Equation
https://www.princeton.edu/~asmits/Bicycle_web/Bernoulli.html
... can give great insight into the balance **between pressure**, **velocity** and elevation. ... When streamlines are parallel the **pressure** is constant across them, except ...

Pressure Vs velocity | Student Doctor Network
forums.studentdoctor.net › ... › MCAT Study Question Q&A
Jul 21, 2009 - 8 posts - 3 authors
Velocity increases with a decrease in pressure. Velocity... ... If you want to think of the **relationship between pressure and velocity**, you can use ...

Genres of Summary

- Extract vs. Abstract

  ...lists fragments of text vs. re-phrases content coherently.

- Single document vs. Multi-document

  ...based on one text vs. fuses together many texts.

- Generic vs. Query-focused

  ...provides author's view vs. reflects user's interest.

Automatic Text Summarization

Genres of Summary

- Extract vs. Abstract

  ...lists fragments of text vs. re-phrases content coherently.

- Single document vs. Multi-document

  ...based on one text vs. fuses together many texts.

- Generic vs. Query-focused

  ...provides author's view vs. reflects user's interest.

Query-focused summarization can be thought of as a complex question answering system

Summarization: Main stages

Content Selection

Choose sentences to extract from the document

Summarization: Main stages

Content Selection

Choose sentences to extract from the document

Information Ordering

Choose an order to place them in the summary

Summarization: Main stages

Content Selection

Choose sentences to extract from the document

Information Ordering

Choose an order to place them in the summary

Sentence realization

Simplify the sentences

Summarization: Main stages

Content Selection

Choose sentences to extract from the document

Information Ordering

Choose an order to place them in the summary

Sentence realization

Simplify the sentences

Removing Redundancy

Increase diversification by removing redundant sentences

Summarization: Main stages

Content Selection

Choose sentences to extract from the document

Information Ordering

Choose an order to place them in the summary

Sentence realization

Simplify the sentences

Removing Redundancy

Increase diversification by removing redundant sentences

The most basic algorithm only does the first stage, *content selection*.

Unsupervised content selection; Luhn (1958)

Intuition

Choose sentences that have salient or informative words

Unsupervised content selection; Luhn (1958)

Intuition

Choose sentences that have salient or informative words

Two approaches to define salient words

- *tf-idf:* weigh each word w_i in document j by tf-idf

$$weight(w_i) = tf_{ij} \times idf_i$$

- *Topic signatures:* choose a smaller set of salient words, specific to that domain

$$weight(w_i) = 1 \text{ if } w_i \text{ is a specific term (use mutual information)}$$

Unsupervised content selection; Luhn (1958)

Intuition

Choose sentences that have salient or informative words

Two approaches to define salient words

- *tf-idf:* weigh each word w_i in document j by tf-idf

$$weight(w_i) = tf_{ij} \times idf_i$$

- *Topic signatures:* choose a smaller set of salient words, specific to that domain

$$weight(w_i) = 1 \text{ if } w_i \text{ is a specific term (use mutual information)}$$

Weighing a sentence

$$weight(s) = \frac{1}{|S|} \sum_{w \in S} weight(w)$$

LexRank: A Graph-based approach

Text Document

Computation is a process following a well defined model ...
A computation can be seen as a purely physical phenomena ...
...

processing →

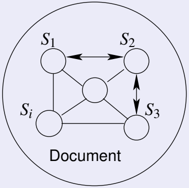
$S_1 \rightarrow \{(computation, 0.1), (process, 0.15), \ldots\}$
$S_2 \rightarrow \{(computation, 0.1), (seen, 0.05), \ldots\}$
$S_3 \rightarrow \ldots$

Machine-readable format

Document Representation

Underlying Hypothesis

Sentences that convey the theme of the document are more similar to each other

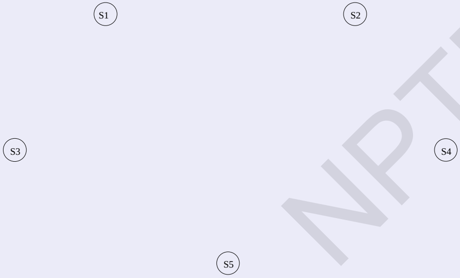


Finding the most salient sentences

Sentence Centrality Measure

Finding the most salient sentences

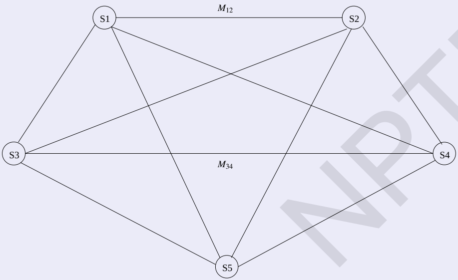
A document graph is constructed with sentences as the vertices



Sentence Centrality Measure

Finding the most salient sentences

A sentence similarity function is used to calculate the edge weights.

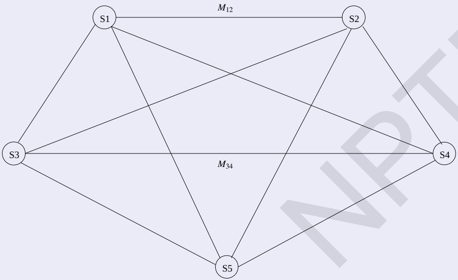


$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

Sentence Centrality Measure

Finding the most salient sentences

PageRank based algorithm is used to compute the sentence centrality vector I.



$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

$$I_j = \mu \cdot \sum_{\forall k \neq j} I_k \cdot \tilde{M}_{k,j} + \frac{1-\mu}{|S|}$$

$$I = \begin{bmatrix} 0.22 & 0.18 & 0.2 & 0.3 & 0.1 \end{bmatrix}$$

Removing Redundant Sentences

Maximal Marginal Relevance

- An iterative method for content selection from a selected list of important sentences
- Iteratively choose the best sentence to insert in the summary that is minimally redundant with the summary so far (Sum)

$$Inf(s)_{MMR} = max_{s \in D}(Inf(s) - \lambda \cdot sim(s, Sum))$$

where $Inf(s)$ denotes the informativeness score of a sentence

Optimization Based Approaches for Summarization

Pawan Goyal

CSE, IIT Kharagpur

Week 11, Lecture 2

Global Inference

- Let us define document D with t_n textual units

$$D = t_1, t_2, \ldots, t_{n-1}, t_n$$

- Let $Rel(i)$ be the relevance of t_i to be in the summary
- Let $Red(i,j)$ be the redundancy between t_i and t_j
- Let $l(i)$ be the length of t_i

Inference Problem

- The inference problem is to select a subset S of textual units from D such that summary score of S, i.e., $s(S)$, is maximized.
- $S = \arg\max_{S \subseteq D} \left[\sum_{t_i \in S} Rel(i) - \sum_{t_i, t_j \in S, i<j} Red(i,j) \right]$
  such that $\sum_{t_i \in S} l(i) \leq K$, where k denotes the maximum length of the summary

A Greedy Solution

1. Sort D so that $Rel(i) > Rel(i+1) \forall i$
2. $S = \{t_1\}$
3. while $\sum_{t_i \in S} l(i) < K$
4. $\quad t_j = \arg\max_{t_j \in D-S} s(S \cup \{t_j\})$
5. $\quad S = S \cup \{t_j\}$
6. return S

Integer Linear Programming (ILP)

- Greedy algorithm is an approximate solution
- Use exact solution algorithms with ILP
- ILP is a constrained optimization problem
- Many solvers on the web
- Define the constraints based on relevance and redundancy for summarization

Sentence Level ILP Formulation

Optimization Function

maximize $\sum_i \alpha_i Rel(i) - \sum_{i<j} \alpha_{ij} Red(i,j)$

Sentence Level ILP Formulation

Optimization Function

maximize $\sum_i \alpha_i Rel(i) - \sum_{i<j} \alpha_{ij} Red(i,j)$

Constraints

such that $\forall i,j$:

- $\alpha_i, \alpha_{ij} \in \{0,1\}$
- $\sum_i \alpha_i l(i) \leq K$
- $\alpha_{ij} - \alpha_i \leq 0$
- $\alpha_{ij} - \alpha_j \leq 0$
- $\alpha_i + \alpha_j - \alpha_{ij} \leq 1$

Sentence Level ILP Formulation

Optimization Function

maximize $\sum_i \alpha_i Rel(i) - \sum_{i<j} \alpha_{ij} Red(i,j)$

Constraints

such that $\forall i,j$:

- $\alpha_i, \alpha_{ij} \in \{0,1\}$
- $\sum_i \alpha_i l(i) \leq K$
- $\alpha_{ij} - \alpha_i \leq 0$
- $\alpha_{ij} - \alpha_j \leq 0$
- $\alpha_i + \alpha_j - \alpha_{ij} \leq 1$

Is generic enough

Depending on your task, you can define your own optimization function and constrains.

The next steps: Sentence Ordering

The next steps: Sentence Ordering

Chronological ordering: the simplest method

List the sentences in the order, they appear in the document

The next steps: Sentence Ordering

Chronological ordering: the simplest method

List the sentences in the order, they appear in the document

Coherence

- Choose orderings that make neighboring sentences similar (by cosine)
- Choose orderings in which neighboring sentences discuss the same entity

The next steps: Sentence Ordering

Chronological ordering: the simplest method

List the sentences in the order, they appear in the document

Coherence

- Choose orderings that make neighboring sentences similar (by cosine)
- Choose orderings in which neighboring sentences discuss the same entity

Topical ordering

Learn the ordering of topics in the source documents

The next steps: Simplifying Sentences

Parse sentences, use rules to decide which modifiers to prune

The next steps: Simplifying Sentences

Parse sentences, use rules to decide which modifiers to prune

- **Initial adverbials:** *For example, on the other hand, as a matter of fact, at this point, ...*

The next steps: Simplifying Sentences

Parse sentences, use rules to decide which modifiers to prune

- **Initial adverbials:** *For example, on the other hand, as a matter of fact, at this point, ...*
- **PPs without named entities:** The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases [*PP to a sustainable number*]

The next steps: Simplifying Sentences

Parse sentences, use rules to decide which modifiers to prune

- **Initial adverbials:** *For example, on the other hand, as a matter of fact, at this point, ...*
- **PPs without named entities:** The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases [*PP to a sustainable number*]
- **Attribution clauses:** Rebels agreed to talks with government officials, *international observers said Tuesday*

The next steps: Simplifying Sentences

Parse sentences, use rules to decide which modifiers to prune

- **Initial adverbials:** *For example, on the other hand, as a matter of fact, at this point, ...*
- **PPs without named entities:** The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases [*PP to a sustainable number*]
- **Attribution clauses:** Rebels agreed to talks with government officials, *international observers said Tuesday*
- **Appositives:** Rajan*, 28, an artist who was living at the time in Philadelphia,* found the inspiration in the back of city magazines

Summarization: Evaluation

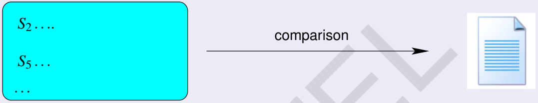
Pawan Goyal

CSE, IIT Kharagpur

Week 11, Lecture 3

System Evaluation

Evaluation Criteria



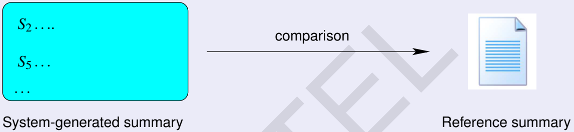
System-generated summary

Reference summary

System Evaluation

System Evaluation

Evaluation Criteria



System Evaluation

ROUGE

Recall Oriented Understudy for Gisting Evaluation *Not as good as human evaluation but much more convenient*

Toolkit available for download.

ROUGE for evaluation

Given a document D, and an automatic summary X:

- Have N humans produce a set of reference summaries of D ($N \geq 1$)
- Run system, giving automatic summary X
- What percentage of the n-grams from the reference summaries appear in X?

$$ROUGE-2 = \frac{\sum_{S \in \{RefSums\}} \sum_{bi-gram \in S} Count_{match}(bi-gram)}{\sum_{S \in \{RefSums\}} \sum_{bi-gram \in S} Count(bi-gram)}$$

ROUGE Example

Reference Summaries

- **Human 1:** water spinach is a green leafy vegetable grown in the tropics.
- **Human 2:** water spinach is a semi-aquatic tropical plant grown as a vegetable.
- **Human 3:** water spinach is a commonly eaten leaf vegetable of Asia

System Summary

water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

ROUGE Example

Reference Summaries

- **Human 1:** water spinach is a green leafy vegetable grown in the tropics.
- **Human 2:** water spinach is a semi-aquatic tropical plant grown as a vegetable.
- **Human 3:** water spinach is a commonly eaten leaf vegetable of Asia

System Summary

water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

ROUGE-2

$$\frac{3+3+6}{10+10+9} = 12/29 = 0.413$$

Further Discussions

Further Discussions

- Multi-document summarization

Further Discussions

- Multi-document summarization
- Query-specific summarization

Further Discussions

- Multi-document summarization
- Query-specific summarization
- Abstractive summarization

Text Classification - I

Pawan Goyal

CSE, IIT Kharagpur

Week 11, Lecture 4

Example: 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.

Example: Male or Female Author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam...
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

Example: What is the subject of this article?

MEDLINE Article

Brain Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Taxt Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language identification
- Sentiment analysis
- ...

Text classification: problem definition

Input

- A document d
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$

Text classification: problem definition

Input

- A document d
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$

Output

A predicted class $c \in C$

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features

Spam

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features

Spam

black-list-address OR (“dollars” AND “have been selected”)

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features

Spam

black-list-address OR (“dollars” AND “have been selected”)

Pros and Cons

Accuracy can be high if rules carefully refined by expert, *but building and maintaining these rules is expensive.*

Classification Methods: Supervised Machine Learning

- Naïve Bayes
- Logistic regression
- Support-vector machines
- ...

Naïve Bayes Intuition

- Simple classification method based on Bayes' rule
- Relies on very simple representation of document - Bag of words

Bag of words for document classification

Test document

parser
language
label
translation
...

?

Machine Learning	NLP	Garbage Collection	Planning	GUI
learning	parser	garbage	planning	...
training	tag	collection	temporal	
algorithm	training	memory	reasoning	
shrinkage	translation	optimization	plan	
network...	language...	region...	language...	

Bayes' rule for documents and classes

For a document d and a class c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Bayes' rule for documents and classes

For a document d and a class c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naïve Bayes Classifier

$$c_{MAP} = \arg\max_{c \in C} P(c|d)$$

$$= \arg\max_{c \in C} P(d|c)P(c)$$

$$= \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n|c)P(c)$$

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Conditional Independence

Assume the feature probabilities $P(x_i | c_j)$ are independent given the class c_j.

$$P(x_1, x_2, \ldots, x_n | c) = P(x_1 | c) \cdot P(x_2 | c) \ldots P(x_n | c)$$

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Conditional Independence

Assume the feature probabilities $P(x_i|c_j)$ are independent given the class c_j.

$$P(x_1, x_2, \ldots, x_n | c) = P(x_1|c) \cdot P(x_2|c) \ldots P(x_n|c)$$

$$c_{NB} = \underset{c \in C}{\arg\max}\, P(c) \prod_{x \in X} P(x|c)$$

Learning the model parameters

Maximum Likelihood Estimate

$$\hat{P}(c_j) = \frac{doc - count(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Learning the model parameters

Maximum Likelihood Estimate

$$\hat{P}(c_j) = \frac{doc - count(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Problem with MLE

Suppose in the training data, we haven't seen the word "fantastic", classified in the topic 'positive'.

$$\hat{P}(fantastic|positive) = 0$$

Learning the model parameters

Maximum Likelihood Estimate

$$\hat{P}(c_j) = \frac{doc - count(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Problem with MLE

Suppose in the training data, we haven't seen the word "fantastic", classified in the topic 'positive'.

$$\hat{P}(fantastic|positive) = 0$$

$$c_{NB} = \arg\max_{c} \hat{P}(c) \prod_{x \in X} \hat{P}(x_i|c)$$

Laplace (add-1) smoothing

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w \in V}(count(w,c)+1)}$$

$$= \frac{count(w_i,c)+1}{(\sum_{w \in V}(count(w,c)) + |V|}$$

Text Classification - II

Pawan Goyal

CSE, IIT Kharagpur

Week 11, Lecture 5

A worked example

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

A worked example

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c)=$

$P(j)=$

A worked example

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|c) =

P(Tokyo|c) =

P(Japan|c) =

P(Chinese|j) =

P(Tokyo|j) =

P(Japan|j) =

A worked example

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7
P(Tokyo|*c*) = (0+1) / (8+6) = 1/14
P(Japan|*c*) = (0+1) / (8+6) = 1/14
P(Chinese|*j*) = (1+1) / (3+6) = 2/9
P(Tokyo|*j*) = (1+1) / (3+6) = 2/9
P(Japan|*j*) = (1+1) / (3+6) = 2/9

A worked example

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Choosing a class:

$P(c|d5) \propto$

$P(j|d5) \propto$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

A worked example

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Choosing a class:

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

Naïve Bayes and Language Modeling

In general, NB classifier can use any feature

URL, email addresses, dictionaries, network features

Naïve Bayes and Language Modeling

In general, NB classifier can use any feature

URL, email addresses, dictionaries, network features

But if we use only the word features and all the words in the text

Naïve Bayes has an important similarity to language modeling.

Naïve Bayes and Language Modeling

In general, NB classifier can use any feature

URL, email addresses, dictionaries, network features

But if we use only the word features and all the words in the text

Naïve Bayes has an important similarity to language modeling.

Each class can be thought of as a separate unigram language model.

Which class assigns a higher probability to the sentence?

Model pos	
0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Model neg	
0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

I	love	this	fun	film
0.1	0.1	0.01	0.05	0.1
0.2	0.001	0.01	0.005	0.1

$$P(s|pos) > P(s|neg)$$

Naïve Bayes: More than Two Classes

Multi-value classification

A document can belong to 0, 1 or > 1 classes

Naïve Bayes: More than Two Classes

Multi-value classification

A document can belong to 0, 1 or > 1 classes

Handling Multi-value classification

- For each class $c \in C$, build a classifier γ_c to distinguish c from all other classes $c' \in C$

Naïve Bayes: More than Two Classes

Multi-value classification

A document can belong to 0, 1 or > 1 classes

Handling Multi-value classification

- For each class $c \in C$, build a classifier γ_c to distinguish c from all other classes $c' \in C$
- Given test-doc d, evaluate it for membership in each class using each γ_c

Naïve Bayes: More than Two Classes

Multi-value classification

A document can belong to 0, 1 or > 1 classes

Handling Multi-value classification

- For each class $c \in C$, build a classifier γ_c to distinguish c from all other classes $c' \in C$
- Given test-doc d, evaluate it for membership in each class using each γ_c
- d belongs to any class for which γ_c returns true

Naïve Bayes: More than Two Classes

One-of or multinomial classification

Classes are mutually exclusive: each document in exactly one class

Naïve Bayes: More than Two Classes

One-of or multinomial classification

Classes are mutually exclusive: each document in exactly one class

Binary classifiers may also be used

- For each class $c \in C$, build a classifier γ_c to distinguish c from all other classes $c' \in C$

Naïve Bayes: More than Two Classes

One-of or multinomial classification

Classes are mutually exclusive: each document in exactly one class

Binary classifiers may also be used

- For each class $c \in C$, build a classifier γ_c to distinguish c from all other classes $c' \in C$
- Given test-doc d, evaluate it for membership in each class using each γ_c

Naïve Bayes: More than Two Classes

One-of or multinomial classification

Classes are mutually exclusive: each document in exactly one class

Binary classifiers may also be used

- For each class $c \in C$, build a classifier γ_c to distinguish c from all other classes $c' \in C$
- Given test-doc d, evaluate it for membership in each class using each γ_c
- d belongs to one class with maximum score

Evaluation: Constructing Confusion matrix c

For each pair of classes $< c_1, c_2 >$ how many documents from c_1 were incorrectly assigned to c_2? (when $c_2 \neq c_1$)

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

Per class evaluation measures

Recall

Per class evaluation measures

Recall

Fraction of docs in class i classified correctly: $\frac{c_{ii}}{\sum_j c_{ij}}$

Per class evaluation measures

Recall

Fraction of docs in class i classified correctly: $\frac{c_{ii}}{\sum_j c_{ij}}$

Precision

Fraction of docs assigned class i that are actually about class i:

Per class evaluation measures

Recall

Fraction of docs in class i classified correctly: $\frac{c_{ii}}{\sum_j c_{ij}}$

Precision

Fraction of docs assigned class i that are actually about class i: $\frac{c_{ii}}{\sum_i c_{ji}}$

Per class evaluation measures

Recall

Fraction of docs in class i classified correctly: $\frac{c_{ii}}{\sum_j c_{ij}}$

Precision

Fraction of docs assigned class i that are actually about class i: $\frac{c_{ii}}{\sum_i c_{ji}}$

Accuracy

Fraction of docs classified correctly: $\frac{\sum_i c_{ii}}{N}$

Micro- vs. Macro-Average

If we have more than one class, how do we combine multiple performance measures into one quantity?

Micro- vs. Macro-Average

If we have more than one class, how do we combine multiple performance measures into one quantity?

Macro-averaging

Compute performance for each class, then average

Micro- vs. Macro-Average

If we have more than one class, how do we combine multiple performance measures into one quantity?

Macro-averaging

Compute performance for each class, then average

Micro-averaging

Collect decisions for all the classes, compute contingency table, evaluate.

Micro- vs. Macro-Average

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

Micro- vs. Macro-Average

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macro-averaged precision:

Micro- vs. Macro-Average

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macro-averaged precision: $(0.5 + 0.9)/2 = 0.7$
- Micro-averaged precision:

Micro- vs. Macro-Average

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macro-averaged precision: $(0.5 + 0.9)/2 = 0.7$
- Micro-averaged precision: $100/120 = 0.83$

Micro- vs. Macro-Average

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macro-averaged precision: (0.5 + 0.9)/2 = 0.7
- Micro-averaged precision: 100/120 = 0.83

Micro-averaged score is dominated by score on common classes