



# Text Classification and Naïve Bayes

# The Task of Text Classification

Dan Jurafsky



# Is this spam?

**Subject: Important notice!**

**From:** Stanford University <newsforum@stanford.edu>

**Date:** October 28, 2011 12:34:16 PM PDT

**To:** undisclosed-recipients;;

---

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

<http://www.123contactform.com/contact-form-StanfordNew1-236335.html>

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

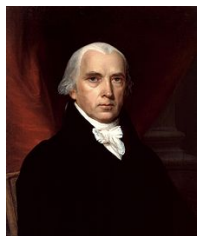
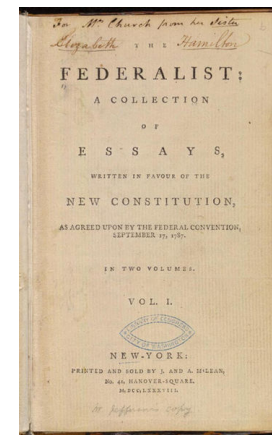
© Stanford University. All Rights Reserved.

Dan Jurafsky

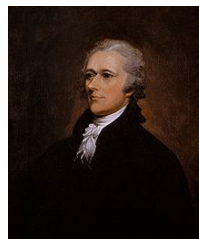


# Who wrote which Federalist papers?

- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods



James Madison



Alexander Hamilton



## 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 Cochinchina; 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...



## Positive or negative movie review?



- unbelievably disappointing



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

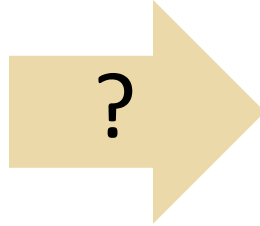


- this is the greatest screwball comedy ever filmed



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

## MEDLINE Article



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



# Text Classification

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



## Text Classification: definition

- *Input*:
  - a document  $d$
  - a fixed set of classes  $C = \{c_1, c_2, \dots, c_J\}$
- *Output*: a predicted class  $c \in C$



Dan Jurafsky



# Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive



# Classification Methods: Supervised Machine Learning

- *Input:*
  - a document  $d$
  - a fixed set of classes  $C = \{c_1, c_2, \dots, c_J\}$
  - A training set of  $m$  hand-labeled documents  $(d_1, c_1), \dots, (d_m, c_m)$
- *Output:*
  - a learned classifier  $\gamma: d \rightarrow c$

Dan Jurafsky



# Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Naïve Bayes
  - Logistic regression
  - Support-vector machines
  - k-Nearest Neighbors
- ...



# Text Classification and Naïve Bayes

# The Task of Text Classification

[illegible][illegible]



## 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	<div><div></div></div>	"This was very easy to setup to four computers."
value	<div><div></div></div>	"Appreciate good quality at a fair price."
setup	<div><div></div></div>	"Overall pretty easy setup."
customer service	<div><div></div></div>	"I DO like honest tech support people."
size	<div><div></div></div>	"Pretty Paper weight."
mode	<div><div></div></div>	"Photos were fair on the high quality mode."
colors	<div><div></div></div>	"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

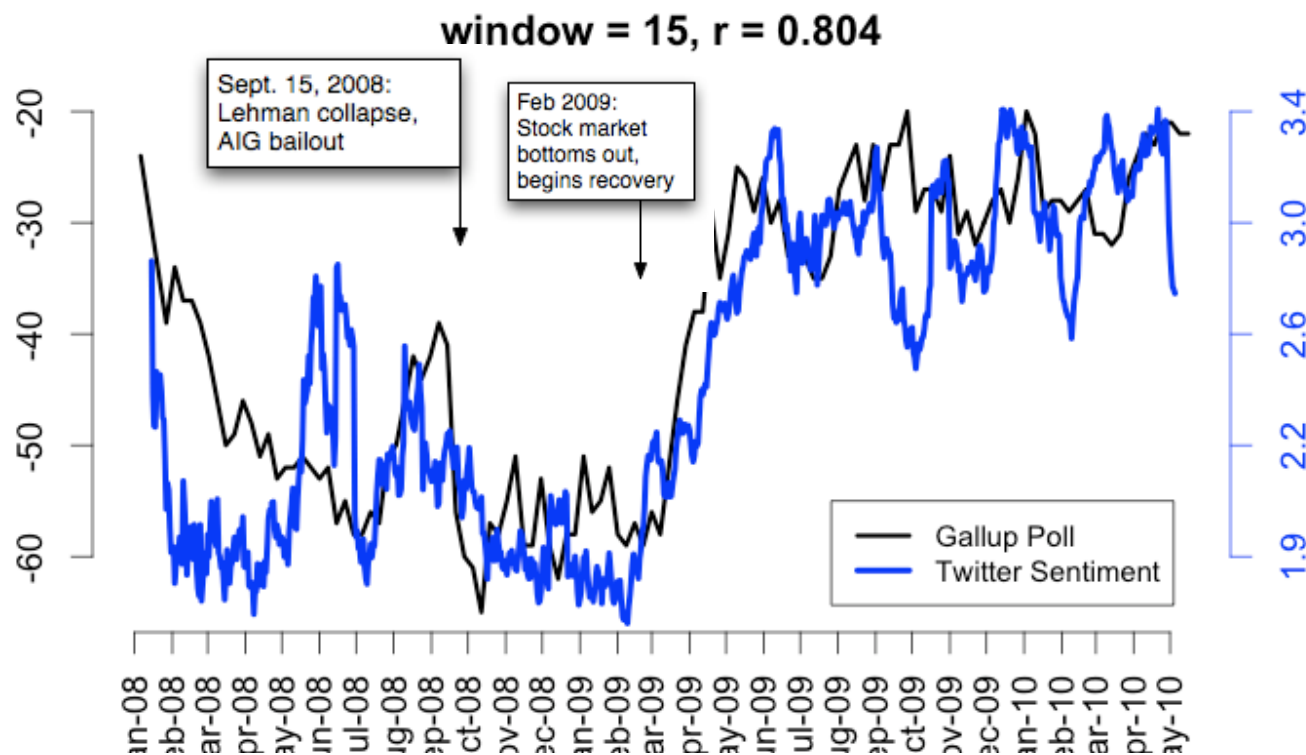
Best Buy (140)  
CNET (5)  
Amazon.com (3)





# Twitter sentiment versus Gallup Poll of Consumer Confidence

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010.  
From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010





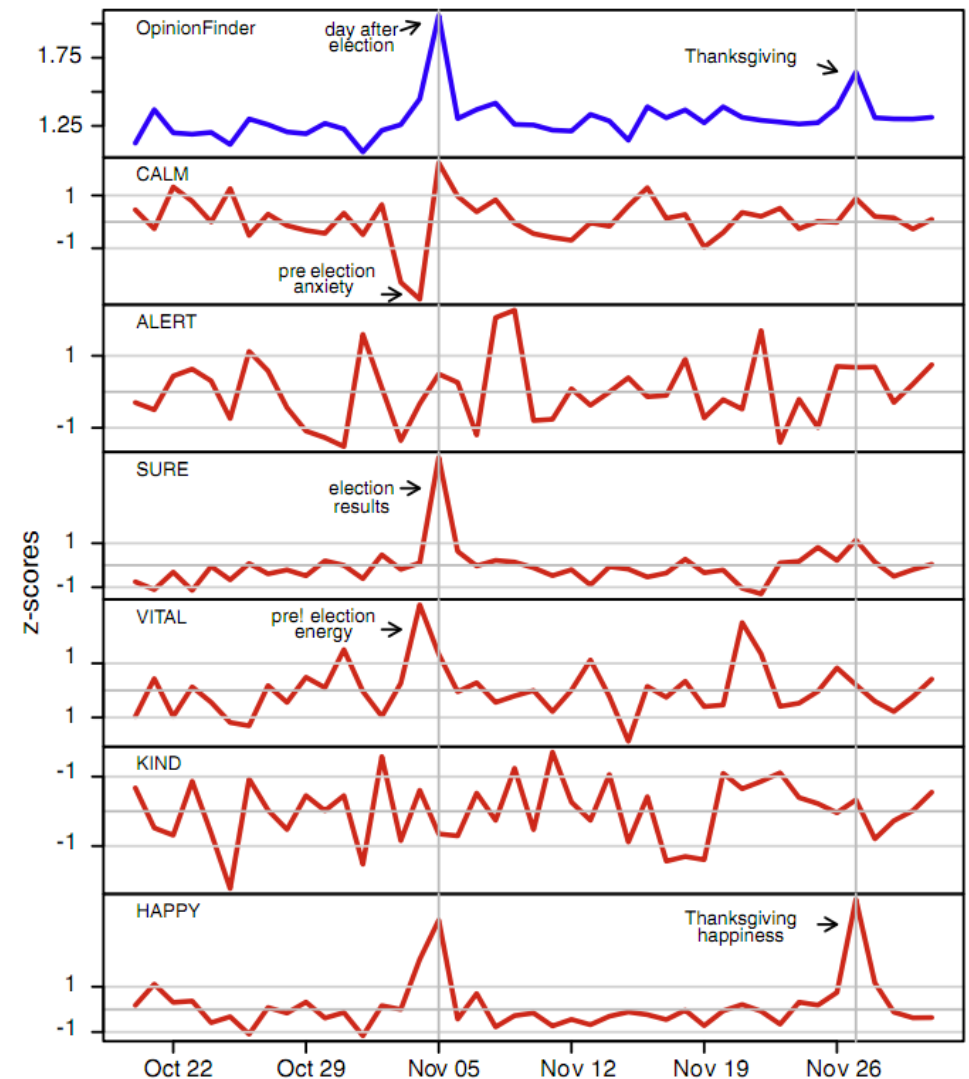
# Twitter sentiment:

Johan Bollen, Huina Mao, Xiaojun Zeng. 2011.

Twitter mood predicts the stock market,

Journal of Computational Science 2:1, 1-8.

10.1016/j.jocs.2010.12.007.

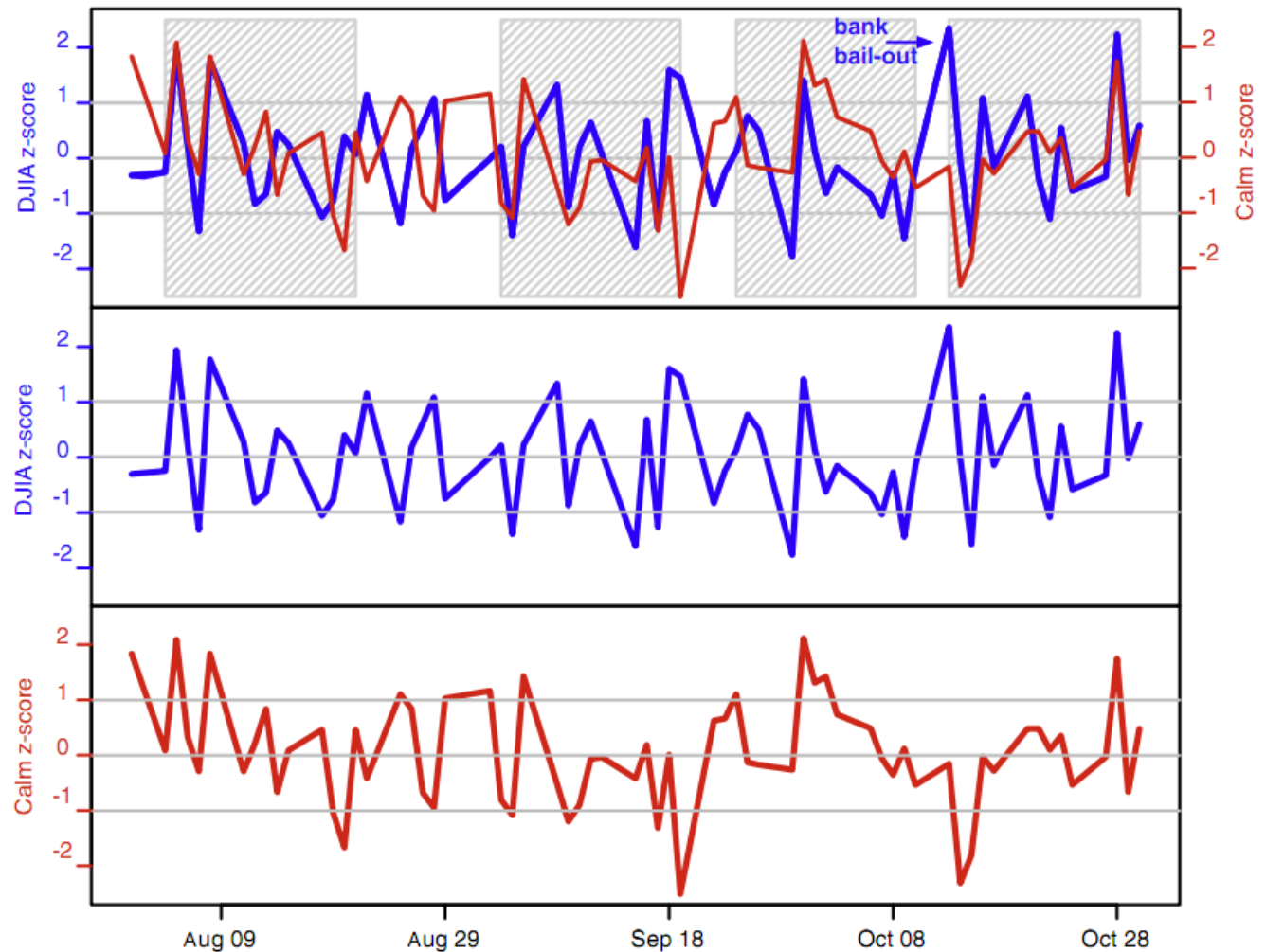




Bollen et al. (2011)

- CALM predicts DJIA 3 days later
- At least one current hedge fund uses this algorithm

CALM Dow Jones





# Target Sentiment on Twitter

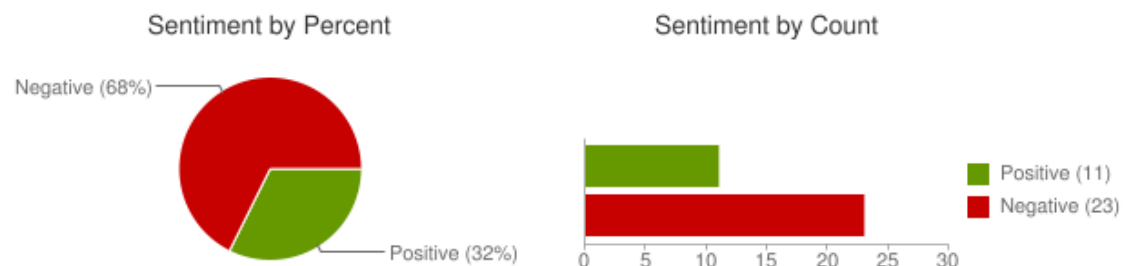
- Twitter Sentiment App

- Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision

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

[Save this search](#)

## Sentiment analysis for "united airlines"



jljacobson: OMG... Could @United airlines have worse customer service? W8g now 15 minut  
Posted 2 hours ago

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

EMLandPRGbelgiu: EML/PRG fly with Q8 united airlines and 24seven to an exotic destination  
Posted 2 hours ago

CountAdam: FANTASTIC customer service from United Airlines at XNA today. Is tweet more,  
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*
- **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*



# 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*
- **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 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



# Question Answering





# Question Answering

One of the oldest NLP tasks (punched card systems in 1961)

Simmons, Klein, McConlogue. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196-204

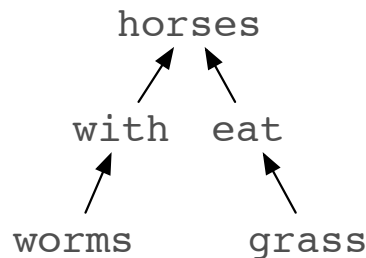
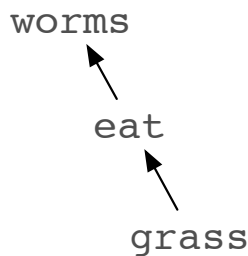
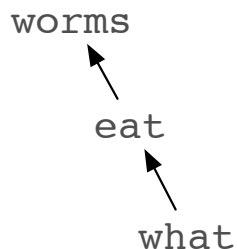
Question:

Potential Answers:

What do worms eat?

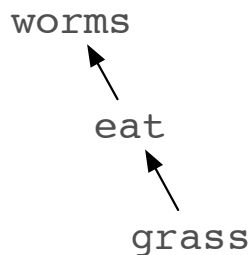
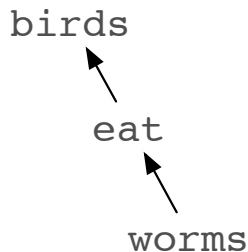
Worms eat grass

Horses with worms eat grass



Birds eat worms

Grass is eaten by worms





# Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S  
"AN ACCOUNT OF THE PRINCIPALITIES OF  
WALLACHIA AND MOLDOVA"  
INSPIRED THIS AUTHOR'S  
MOST FAMOUS NOVEL



Bram Stoker



# Apple's Siri







how many calories are in two slices of banana cream pie?



Examples Random

Assuming any type of pie, banana cream | Use pie, banana cream, prepared from recipe or pie, banana cream, no-bake type, prepared from mix instead

Input interpretation:

pie	amount	2 slices	total calories
	type	banana cream	

Average result:

Show details

702 Cal (dietary Calories)



# Types of Questions in Modern Systems

- Factoid questions
  - *Who wrote “The Universal Declaration of Human Rights”?*
  - *How many calories are there in two slices of apple pie?*
  - *What is the average age of the onset of autism?*
  - *Where is Apple Computer based?*
- Complex (narrative) questions:
  - *In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?*
  - *What do scholars think about Jefferson’s position on dealing with pirates?*



# Commercial systems: mainly factoid questions

Where is the Louvre Museum located?	In Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	The yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What is the telephone number for Stanford University?	650-723-2300



# Paradigms for QA

- IR-based approaches
  - TREC; IBM Watson; Google
- Knowledge-based and Hybrid approaches
  - IBM Watson; Apple Siri; Wolfram Alpha; True Knowledge Evi



# Many questions can already be answered by web search



What are the names of Odin's ravens?

Search

About 214,000 results (0.38 seconds)

Everything

[Huginn and Muninn - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Huginn\\_and\\_Muninn](https://en.wikipedia.org/wiki/Huginn_and_Muninn)

Images

The **names** of the **ravens** are sometimes modernly anglicized as Hugin and Munin. In the Poetic Edda, a disguised **Odin** expresses that he fears that they may ...

Maps

[Attestations](#) - [Archaeological record](#) - [Theories](#) - [See also](#)



# IR-based Question Answering



Where is the Louvre Museum located?

Search

About 904,000 results (0.30 seconds)

Everything

Best guess for Louvre Museum Location is **Paris, France**

Images

Mentioned on at least 7 websites including [wikipedia.org](#), [answers.com](#) and [east-buc.k12.ia.us](#) - [Show sources](#) - [Feedback](#)

Maps

[Musée du Louvre - Wikipedia, the free encyclopedia](#)  
[en.wikipedia.org/wiki/Musée\\_du\\_Louvre](#)

Videos

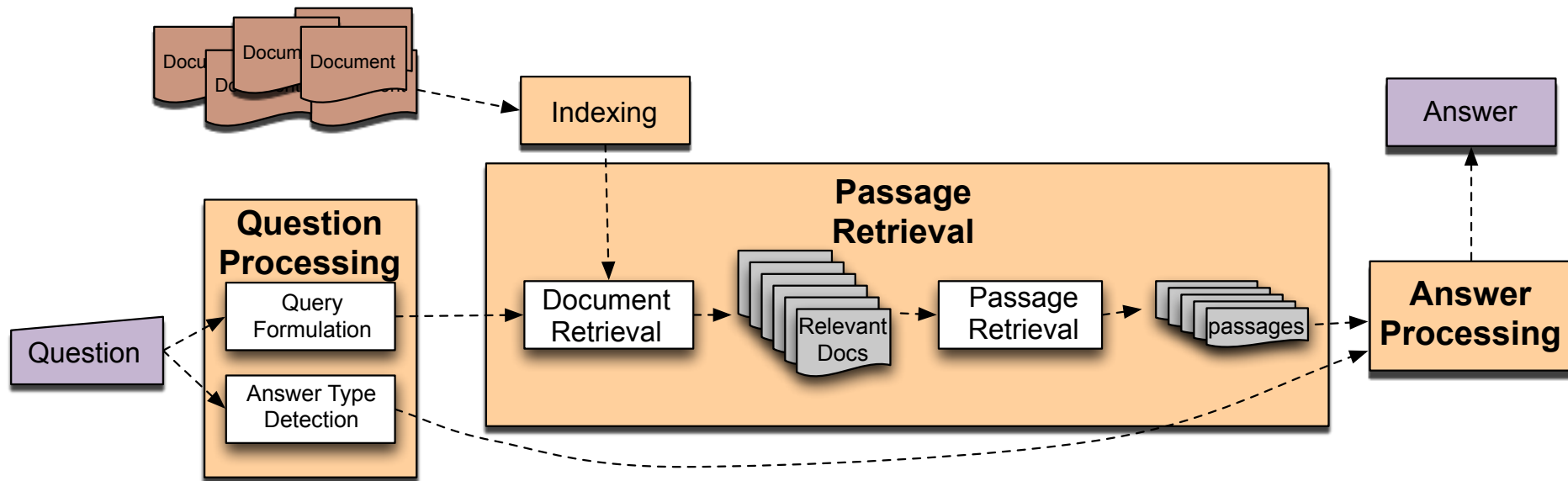
Musée du **Louvre** is **located** in Paris. **Location** within Paris. Established, 1793. **Location**, **Palais Royal**, Musée du **Louvre**, **75001 Paris, France**. Type, Art **museum** ...

News

[Louvre Palace](#) - [List of works in the Louvre](#) - [Category:Musée du Louvre](#)



# IR-based Factoid QA





# IR-based Factoid QA

- **QUESTION PROCESSING**
  - Detect question type, answer type, focus, relations
  - Formulate queries to send to a search engine
- **PASSAGE RETRIEVAL**
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- **ANSWER PROCESSING**
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources





# Knowledge-based approaches (Siri)

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases



# Hybrid approaches (IBM Watson)

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
  - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

# Question Answering



[illegible]

# Summarization in Question Answering



# Text Summarization

- **Goal:** produce an abridged version of a text that contains information that is important or relevant to a user.
- **Summarization Applications**
  - **outlines or abstracts** of any document, article, etc
  - **summaries** of email threads
  - **action items** from a meeting
  - **simplifying** text by compressing sentences



# What to summarize?

## Single vs. multiple documents

- **Single-document summarization**
  - Given a single document, produce
    - abstract
    - outline
    - headline
- **Multiple-document summarization**
  - Given a group of documents, produce a gist of the content:
    - a series of news stories on the same event
    - a set of web pages about some topic or question



# Query-focused Summarization & Generic Summarization

- Generic summarization:
  - Summarize the content of a document
- Query-focused summarization:
  - summarize a document with respect to an information need expressed in a user query.
  - a kind of complex question answering:
    - Answer a question by summarizing a document that has the information to construct the answer



# Summarization for Question Answering: Snippets

- Create **snippets** summarizing a web page for a query
  - Google: 156 characters (about 26 words) plus title and link

The screenshot shows a Google search interface. The search bar contains the text "what is die brücke?". Below the search bar, it says "Search" and "About 5,910,000 results (0.28 seconds)". On the left side, there is a vertical menu with options: "Everything", "Images", "Maps", "Videos", "News", "Shopping", "Applications", and "More". The "Everything" option is selected. The search results are displayed on the right. The first result is titled "Die Brücke - Wikipedia, the free encyclopedia" with a green link to "en.wikipedia.org/wiki/Die\_Brücke". The snippet below the title reads: "Die Brücke (The Bridge) was a group of German expressionist artists formed in Dresden in 1905, after which the Brücke Museum in Berlin was named. Founding ... You've visited this page 5 times. Last visit: 4/16/12". The second result is titled "Die Brücke (film) - Wikipedia, the free encyclopedia" with a green link to "en.wikipedia.org/wiki/Die\_Brücke\_(film)". The snippet below the title reads: "Die Brücke (English: The Bridge) is a 1959 West German film directed by Austrian filmmaker Bernhard Wicki. It is based on the eponymous 1958 novel by ...". The third result is titled "Die Brücke - Die Brücke Art" with a green link to "www.huntfor.com/arthistory/c20th/diebrucke.htm". The snippet below the title reads: "Die Brücke was the association of artist expressionists from Dresden, Germany. ... Die Brücke made use of a technique that was controlled, intentionally ...".





# Summarization for Question Answering: Multiple documents

Create **answers** to complex questions summarizing multiple documents.

- Instead of giving a snippet for each document
- Create a cohesive answer that combines information from each document

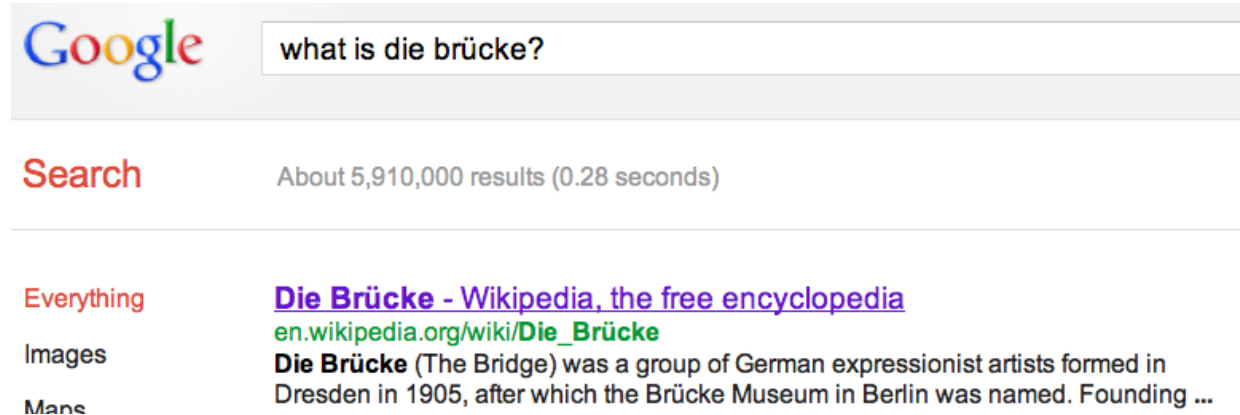


# Extractive summarization & Abstractive summarization

- **Extractive summarization:**
  - create the summary from phrases or sentences in the source document(s)
- **Abstractive summarization:**
  - express the ideas in the source documents using (at least in part) different words



# Simple baseline: take the first sentence



## Die Brücke

From Wikipedia, the free encyclopedia

*For other uses, see [Die Brücke \(disambiguation\)](#).*

**Die Brücke** (The Bridge) was a group of [German expressionist](#) artists formed in [Dresden](#) in 1905, after which the [Brücke Museum in Berlin](#) was named. Founding members were [Fritz Bleyl](#), [Erich Heckel](#), [Ernst Ludwig Kirchner](#) and [Karl Schmidt-Rottluff](#). Later members were [Emil Nolde](#), [Max Pechstein](#) and [Otto Mueller](#). The seminal group had a major impact on the evolution of [modern art](#) in the 20th century and the creation of expressionism.<sup>[1]</sup>

8

Die Brücke is sometimes compared to the [Fauves](#). Both movements shared interests in [primitivist](#) art. Both

[illegible]

# Summarization in Question Answering