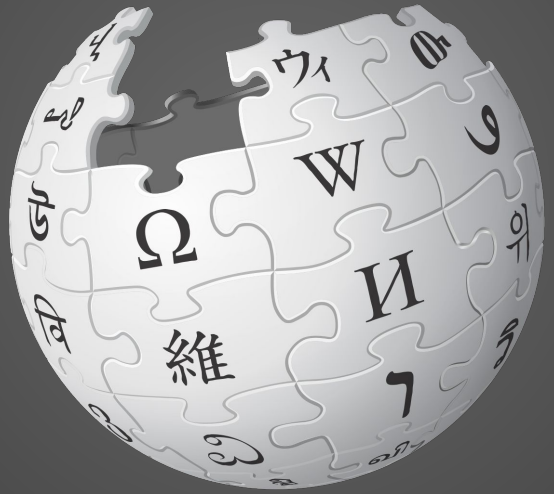
Human Navigation of Information Systems: Wikispeedia

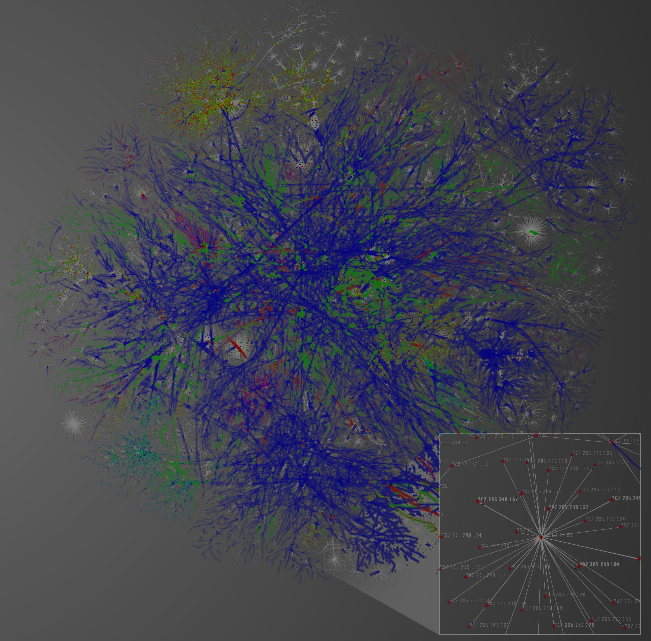


5/5/17 - APAM 4990 - Jake Hofman
Andres, Jayati, and Tomer

Outline

Our goal

- Wikispeedia Data Set
- Exploratory Data Analysis
- Modeling Task
  - Topic Modeling
  - Regression



Motivation

How do humans navigate through information systems?

Motivation

How do humans navigate through information systems?

Can systems be designed to be easier for humans to navigate?

Motivation

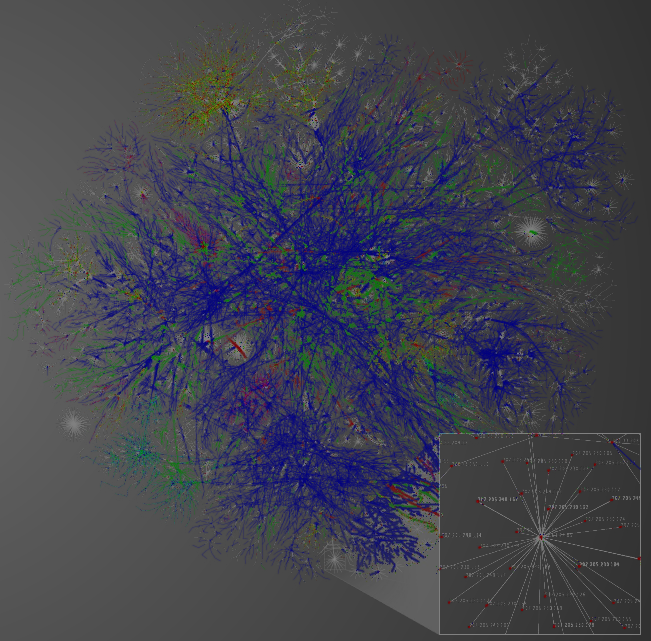
How do humans navigate through information systems?

Can systems be designed to be easier for humans to navigate?

This is especially relevant today!

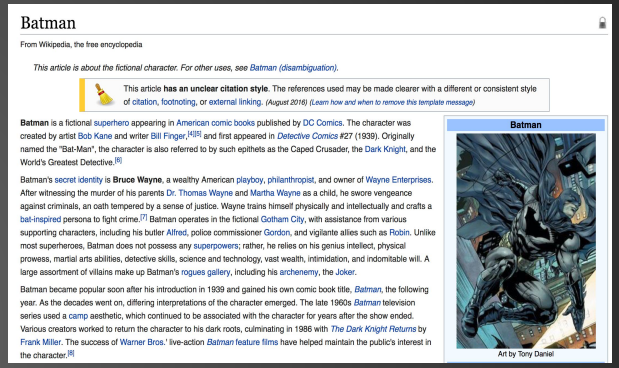
Outline

- Our goal
- **Wikispeedia Data Set**
- Exploratory Data Analysis
- Modeling Task
  - Topic Modeling
  - Regression



How do you play?

- How do you get from Article to A and B with the shortest amount of hyperlinks possible?
- Example
  - Batman => Vitamin D
  - Batman => Superman => Sun => Sunlight => Vitamin D
- Example
  - The Beatles => Beer
  - The Beatles => United States => Household income in the United States=> Oregon => Wine => Beer



Batman

From Wikipedia, the free encyclopedia

This article is about the fictional character. For other uses, see Batman (disambiguation).

This article **has an unclear citation style**. The references used may be made clearer with a different or consistent style of citation, footnoting, or external linking. *(August 2016) (Learn how and when to remove this template message)*

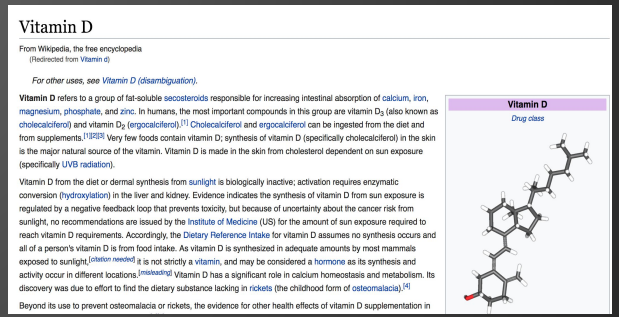
Batman is a fictional superhero appearing in American comic books published by DC Comics. The character was created by artist Bob Kane and writer Bill Finger,[4][5] and first appeared in *Detective Comics* #27 (1939). Originally named the "Bat-Man", the character is also referred to by such epithets as the Caped Crusader, the Dark Knight, and the World's Greatest Detective.[6]

Batman's secret identity is **Bruce Wayne**, a wealthy American playboy, philanthropist, and owner of Wayne Enterprises. After witnessing the murder of his parents Dr. Thomas Wayne and Martha Wayne as a child, he swore vengeance against criminals, an oath tempered by a sense of justice. Wayne trains himself physically and intellectually and crafts a bat-inspired persona to fight crime.[7] Batman operates in the fictional Gotham City, with assistance from various supporting characters, including his butler Alfred, police commissioner Gordon, and vigilante allies such as Robin. Unlike most superheroes, Batman does not possess any superpowers; rather, he relies on his genius intellect, physical prowess, martial arts abilities, detective skills, science and technology, vast wealth, intimidation, and indomitable will. A large assortment of villains make up Batman's rogues gallery, including his archenemy, the Joker.

Batman became popular soon after his introduction in 1939 and gained his own comic book title, *Batman*, the following year. As the decades went on, differing interpretations of the character emerged. The late 1960s *Batman* television series used a camp aesthetic, which continued to be associated with the character for years after the show ended. Various creators worked to return the character to his dark roots, culminating in 1986 with *The Dark Knight Returns* by Frank Miller. The success of Warner Bros.' live-action *Batman* feature films have helped maintain the public's interest in the character.[8]

Batman

Art by Tony Daniel



Vitamin D

From Wikipedia, the free encyclopedia
(Redirected from Vitamin d)

For other uses, see Vitamin D (disambiguation).

Vitamin D refers to a group of fat-soluble secosteroids responsible for increasing intestinal absorption of calcium, iron, magnesium, phosphate, and zinc. In humans, the most important compounds in this group are vitamin D_3 (also known as cholecalciferol) and vitamin D_2 (ergocalciferol).[1] Cholecalciferol and ergocalciferol can be ingested from the diet and from supplements.[1][2][3] Very few foods contain vitamin D; synthesis of vitamin D (specifically cholecalciferol) in the skin is the major natural source of the vitamin. Vitamin D is made in the skin from cholesterol dependent on sun exposure (specifically UVB radiation).

Vitamin D from the diet or dermal synthesis from sunlight is biologically inactive; activation requires enzymatic conversion (hydroxylation) in the liver and kidney. Evidence indicates the synthesis of vitamin D from sun exposure is regulated by a negative feedback loop that prevents toxicity, but because of uncertainty about the cancer risk from sunlight, no recommendations are issued by the Institute of Medicine (US) for the amount of sun exposure required to reach vitamin D requirements. Accordingly, the Dietary Reference Intake for vitamin D assumes no synthesis occurs and all of a person's vitamin D is from food intake. As vitamin D is synthesized in adequate amounts by most mammals exposed to sunlight,[*citation needed*] it is not strictly a vitamin, and may be considered a hormone as its synthesis and activity occur in different locations.[*misleading*] Vitamin D has a significant role in calcium homeostasis and metabolism. Its discovery was due to effort to find the dietary substance lacking in rickets (the childhood form of osteomalacia).[4]

Beyond its use to prevent osteomalacia or rickets, the evidence for other health effects of vitamin D supplementation in

Vitamin D

Drug class

Raw Data

- Source: http://snap.stanford.edu/data/wikispeedia.html
- User paths - Incomplete
- User paths - Complete
- Category mapping for articles
- Links each article contains
- Plaintext files of all the articles

Total number of articles: ~4,600
Total number of games played: ~51,000

Previous Research

WWW 2012 – Session: Web User Behavioral Analysis and Modeling April 16–20, 2012, Lyon, France

Human Wayfinding in Information Networks

Robert West
Computer Science Department
Stanford University
west@cs.stanford.edu

Jure Leskovec
Computer Science Department
Stanford University
jure@cs.stanford.edu

ABSTRACT

Navigating information spaces is an essential part of our everyday lives, and in order to design efficient and user-friendly information systems, it is important to understand how humans navigate and find the information they are looking for. We perform a large-scale study of human wayfinding, in which, given a network of links between the concepts of Wikipedia, people play a game of finding a short path from a given start to a given target concept by following hyperlinks. What distinguishes our setup from other studies of human Web-browsing behavior is that in our case people navigate a graph of connections between concepts, and that the exact goal of the navigation is known ahead of time. We study more than 30,000 goal-directed human search paths and identify strategies people use when navigating information spaces. We find that human wayfinding, while mostly very efficient, differs from shortest paths in characteristic ways. Most subjects navigate through high-degree hubs in the early phase, while their search is guided by content features

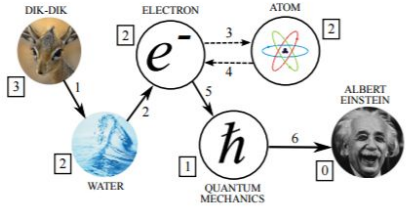
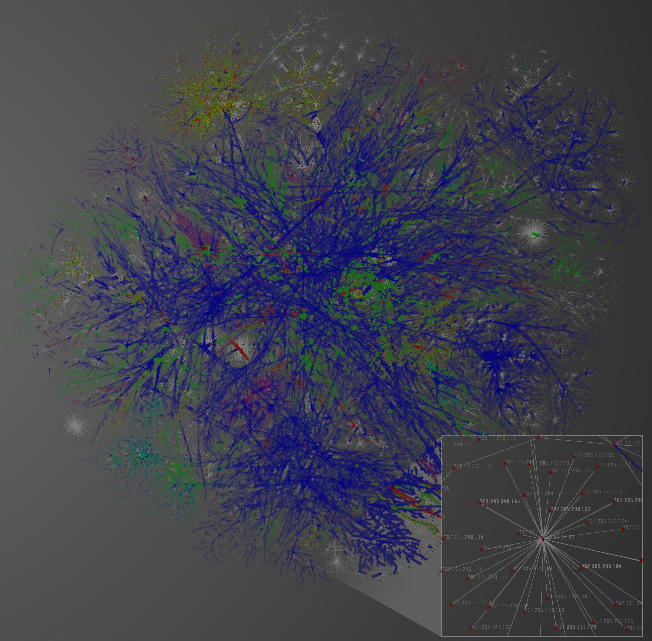


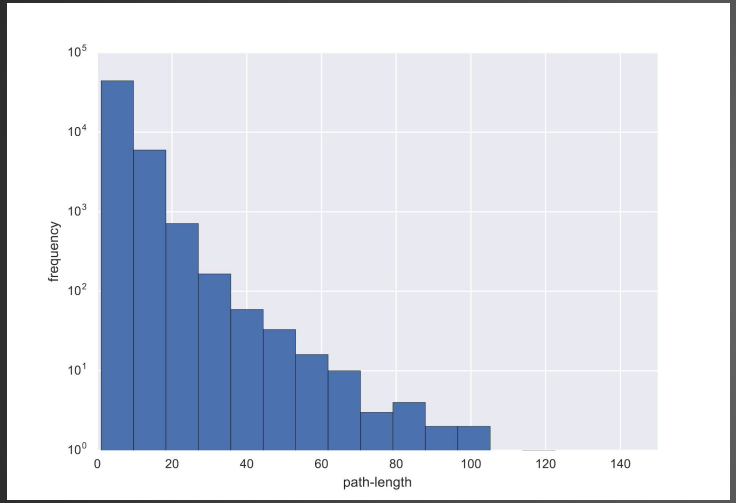
Figure 1: A human example path between the concepts DIK-DIK and ALBERT EINSTEIN. Nodes represent Wikipedia articles and edges the hyperlinks clicked by the human. Edge labels indicate the order of clicks, the framed numbers the shortest-path length to the target. One of several optimal solutions would be ⟨DIK-DIK, WATER, GERMANY, ALBERT EINSTEIN⟩.

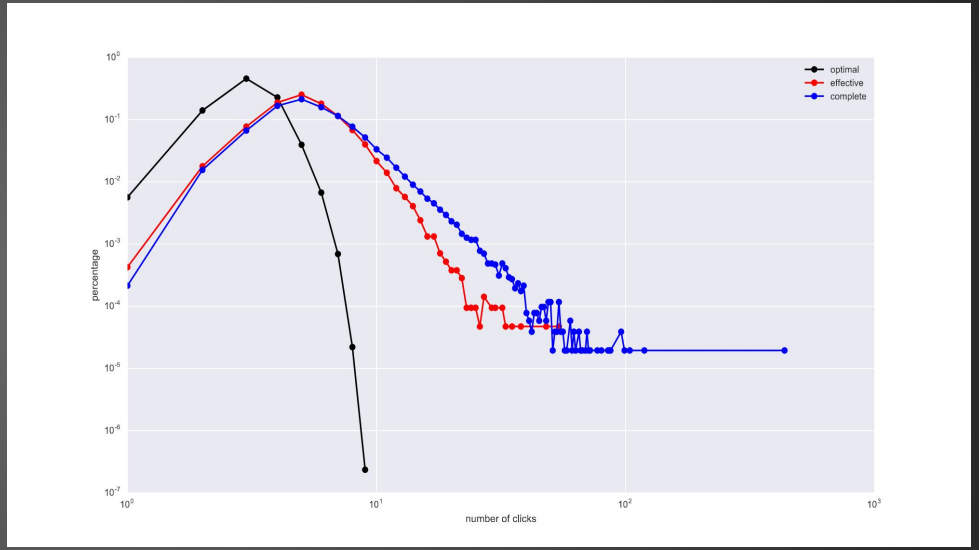
Outline

- Our goal
- Wikispeedia Data Set
- **Exploratory Data Analysis**
- Modeling Task
  - Clustering
  - Regression



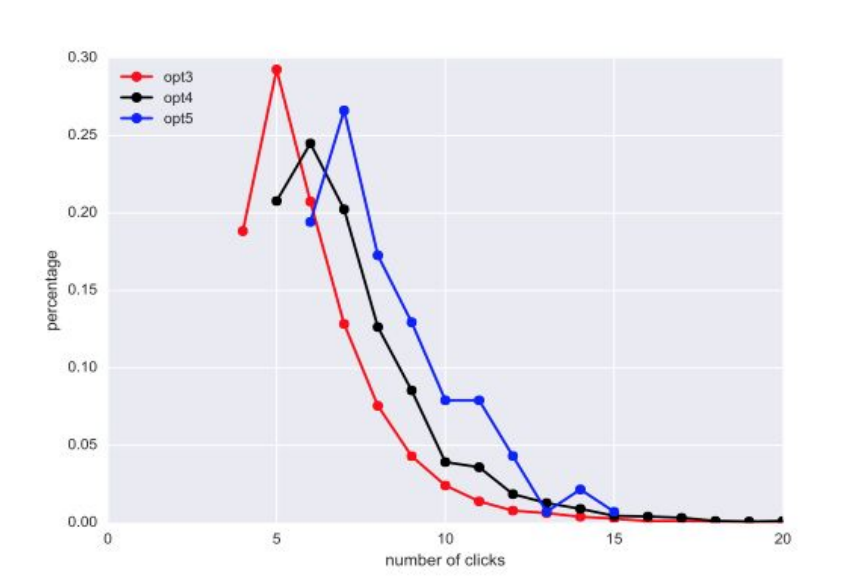
Exploratory Data Analysis: Basic Statistics





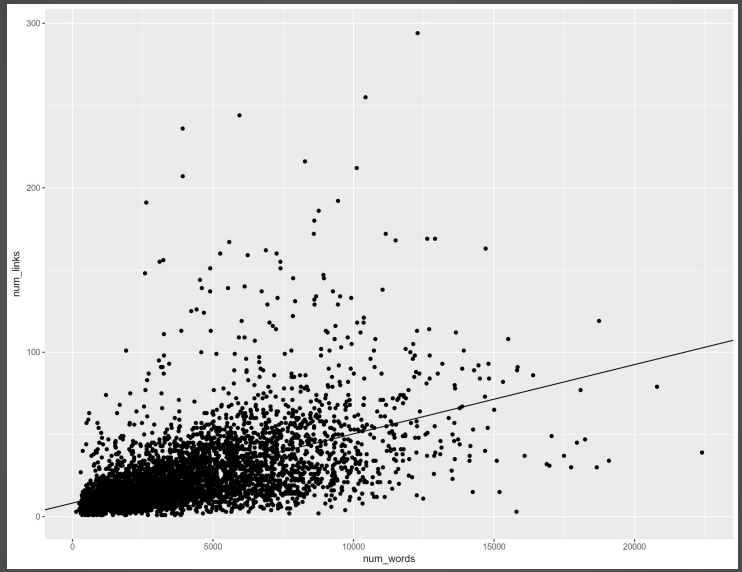
For all games, users tend to do worse than the optimal.

Exploratory Data Analysis: User Performance



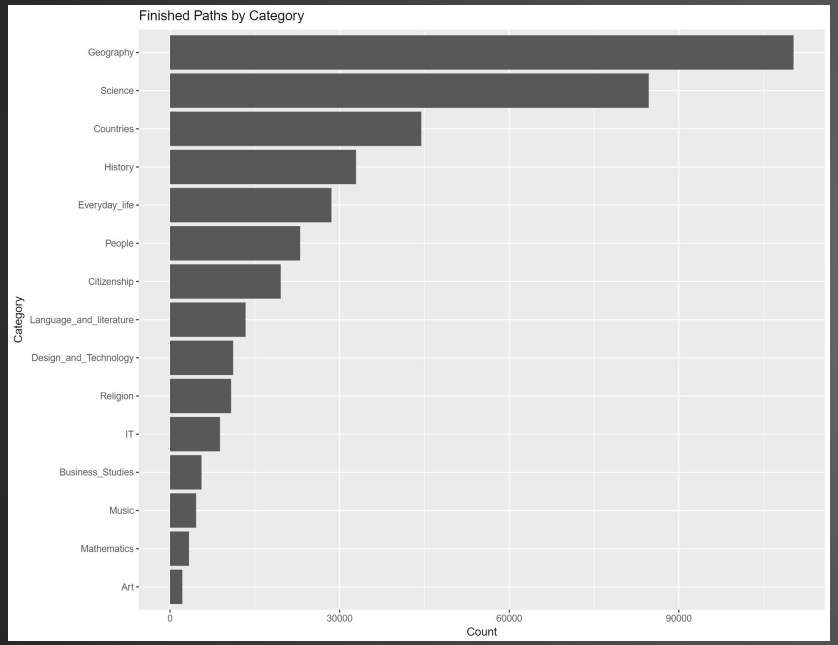
Distribution of optimal paths shows this is a small world network

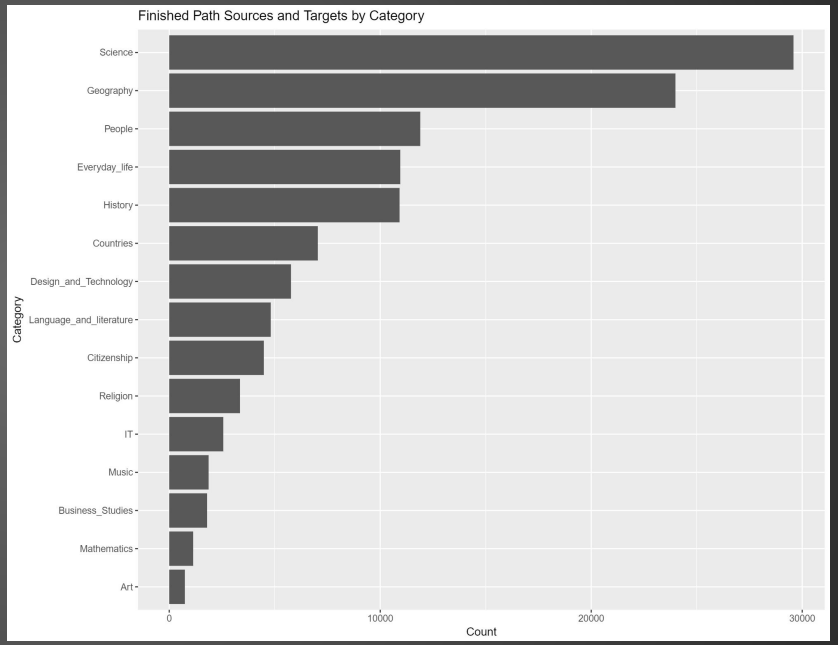
Exploratory Data Analysis: Regression



No significant correlation exists between article size and number of hyperlinks it contains ($R^2 = 0.27$)

Exploratory Data Analysis: Categories

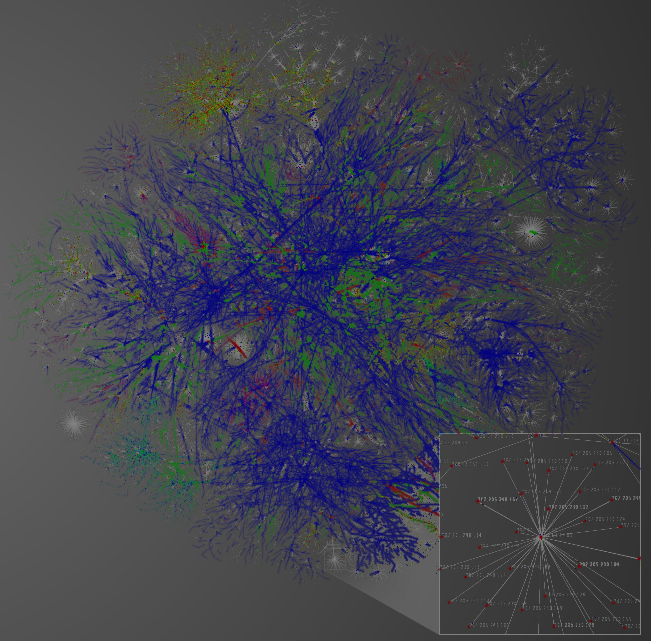




Path category distribution maps well to source/target category distribution

Outline

- Our goal
- Wikispeedia Data Set
- Exploratory Data Analysis
- **Modeling Task**
  - Topic Modeling
  - Regression



Clarifying Our Goal

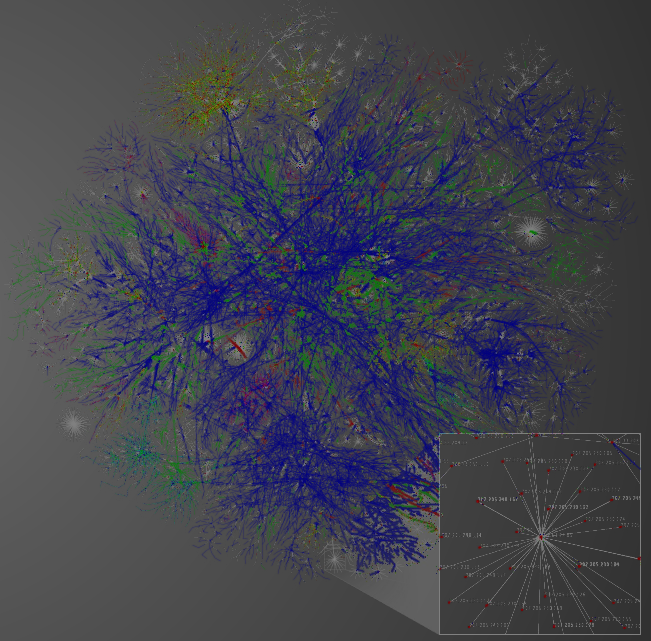
- Humans likely play Wikispeedia by using conceptual connections. However, it may be faster to go from source to target through a non intuitive path.
- **Latent semantic analysis** defines a **semantic distance** between source and target articles to measure the conceptual connection between them.
- **Shortest path length** defines a **topological distance** between source and target articles to measure the optimal path between them.

Can semantic distance be a more effective way to organize Wikipedia categories?

Can we better predict human Wikispeedia performance by using semantic distance rather than topological distance between source and target?

Outline

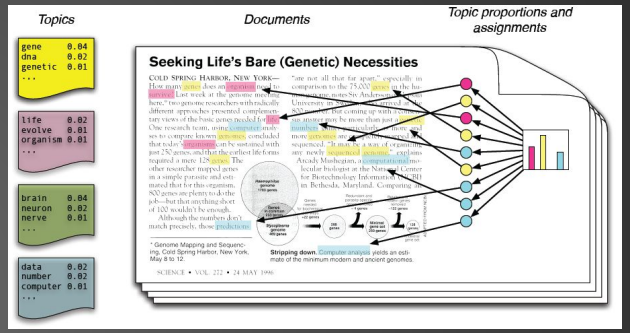
- Our goal
- Wikispeedia Data Set
- Exploratory Data Analysis
- Modeling Task
  - **Topic Modeling**
  - Regression

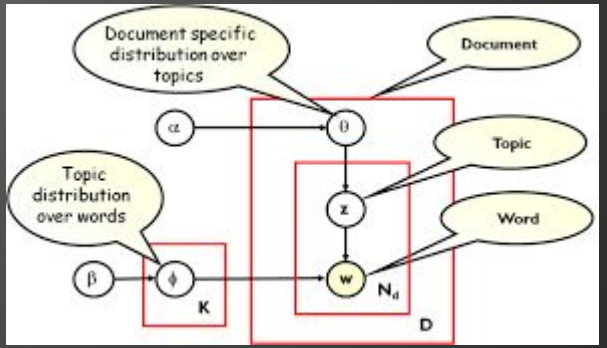


Can semantic distance be a more effective way to organize Wikipedia categories?

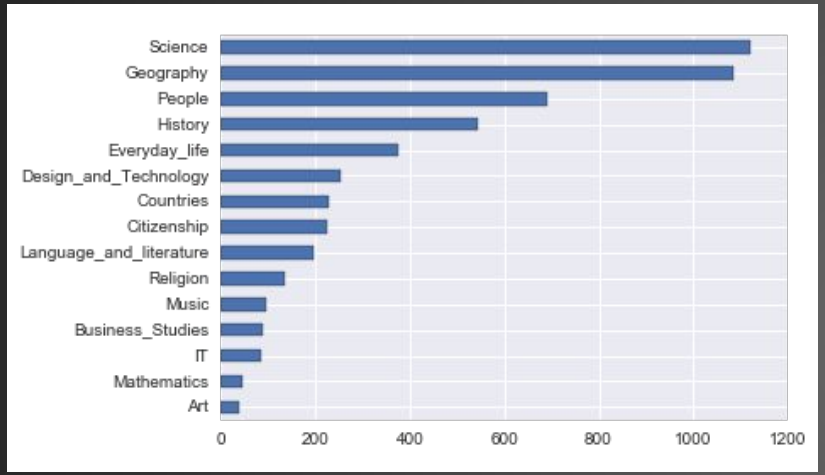
Topic Modeling using Latent Dirichlet Allocation

- Motivation
  - Is there room for improvement in the categories defined by Wikipedia?
  - Higher level, can these categories reveal information about the semantics/clustering of different categories?
- How it works
  - Topic is a distribution over words
  - Document is a mixture of corpus-wide topics
  - Word is drawn from one of those topics
- What we did
  - LDA across all 4,000 articles in dataset with 15 topics

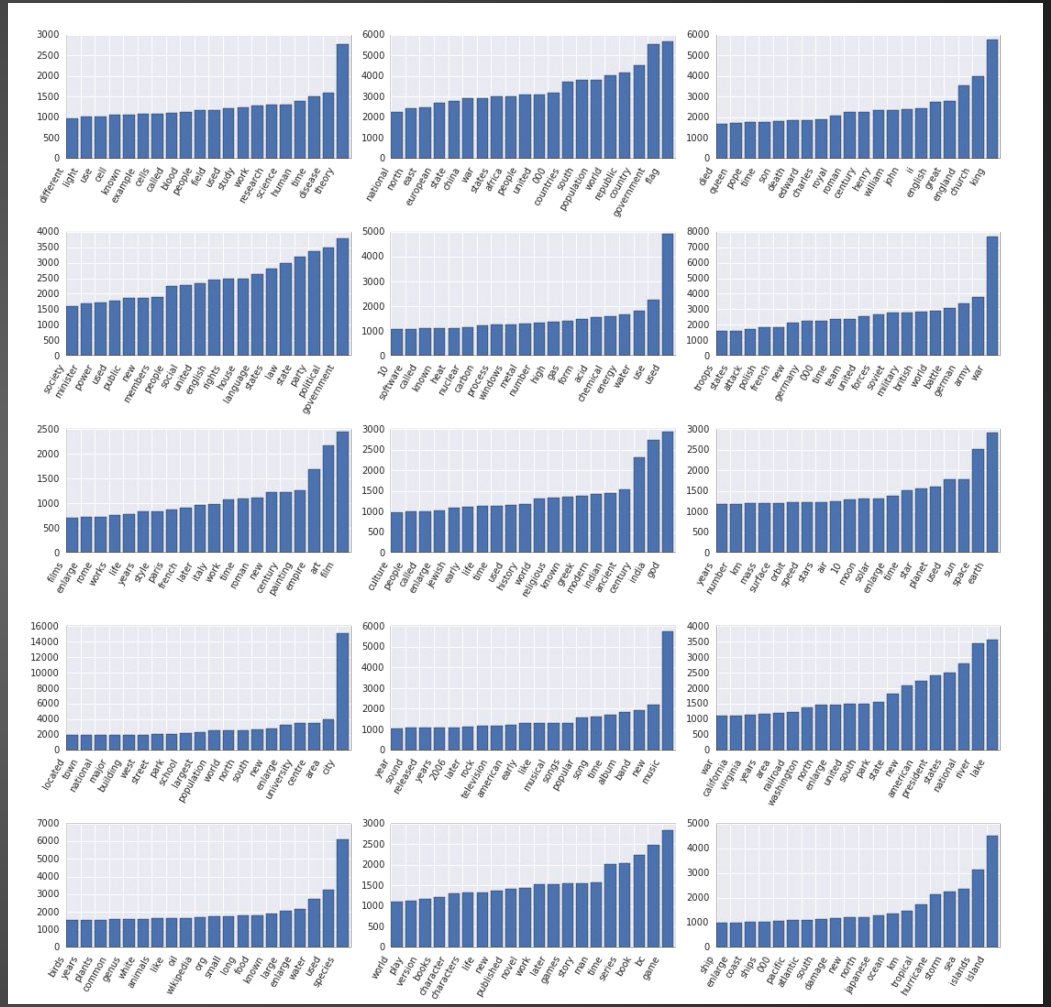




Topic Modeling: Results



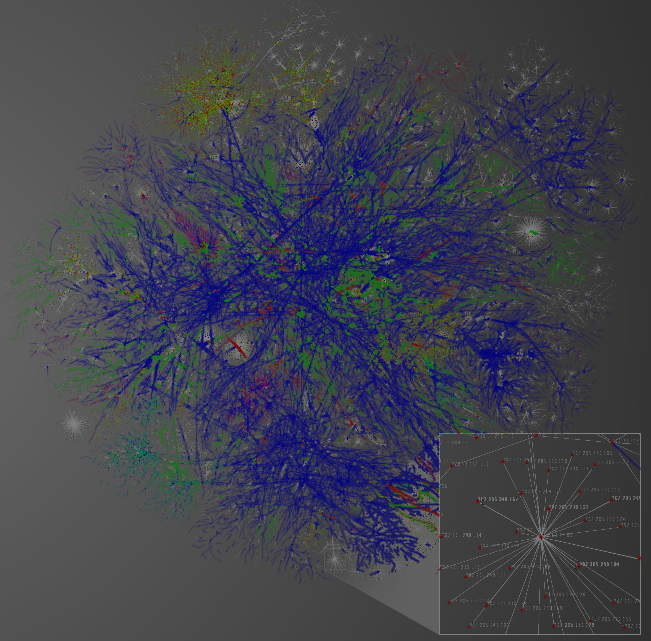
LDA map to categories and provides semantic meaning



Outline

- Our goal
- Wikispeedia Data Set
- Exploratory Data Analysis
- Modeling Task
  - Topic Modeling

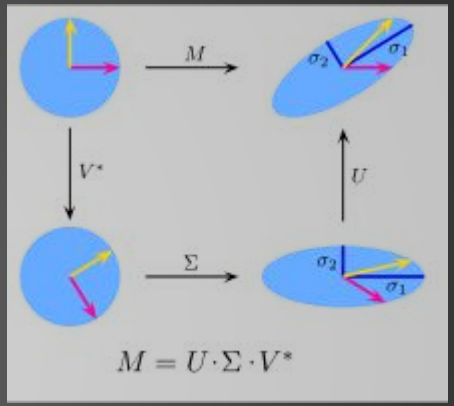
Regression

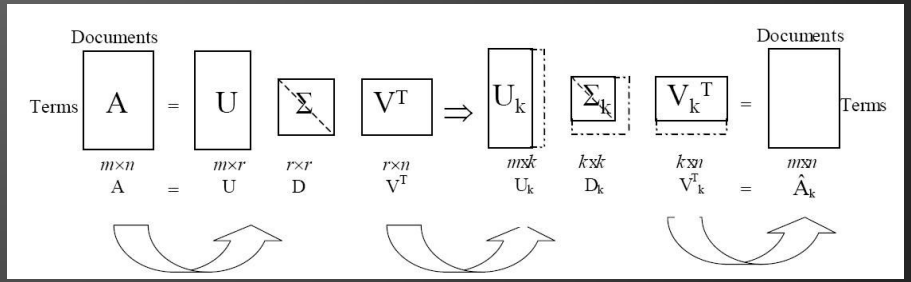


Can we better predict human Wikispeedia performance by using semantic distance rather than topological distance between source and target?

Latent Semantic Analysis to Measure Document Similarity

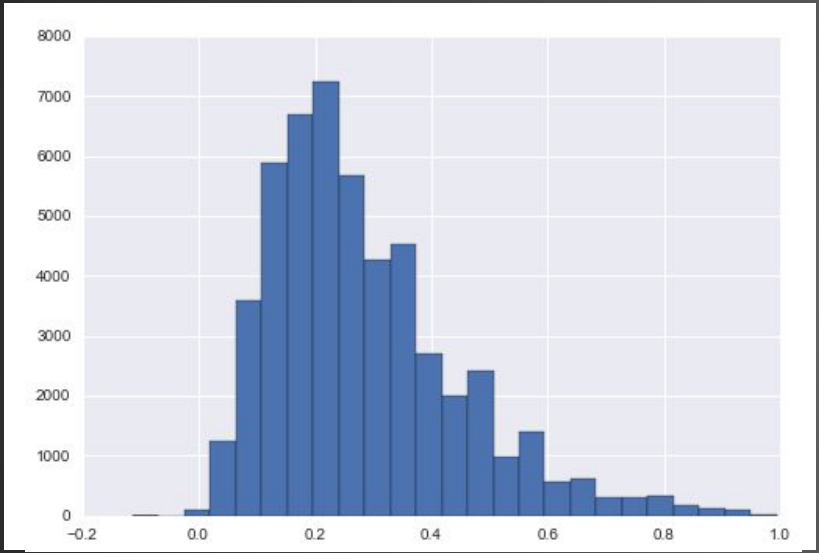
- Motivation
  - Proxy for human intuition of navigating network by going through documents that are related
- How it works
  - Truncated Singular Value Decomposition
  - Dimensionality Reduction Technique
  - Applied to Bag of Words generated from CountVectorizer
- What we did
  - Computed semantic distance between every article using LSA





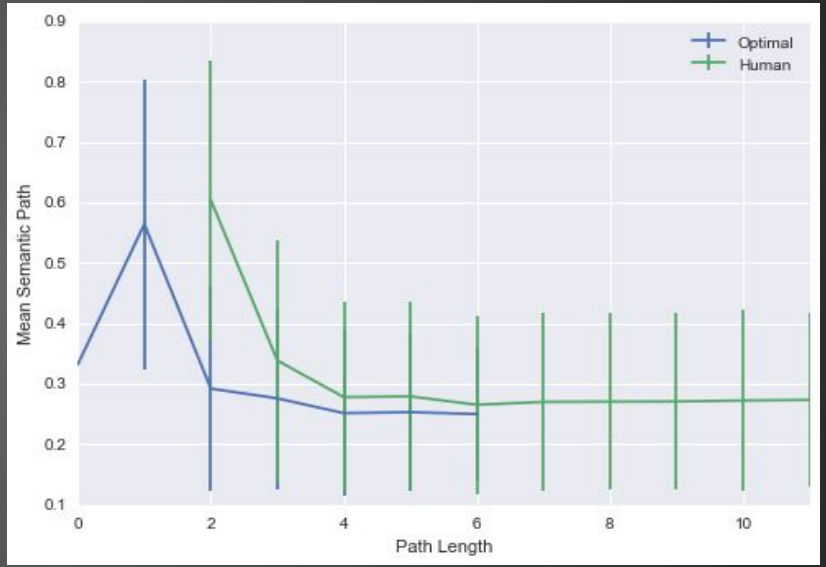
Latent Semantic Analysis: Results

Histogram of Semantic Lengths



Slightly skewed normal distribution

Topological versus Semantic Lengths

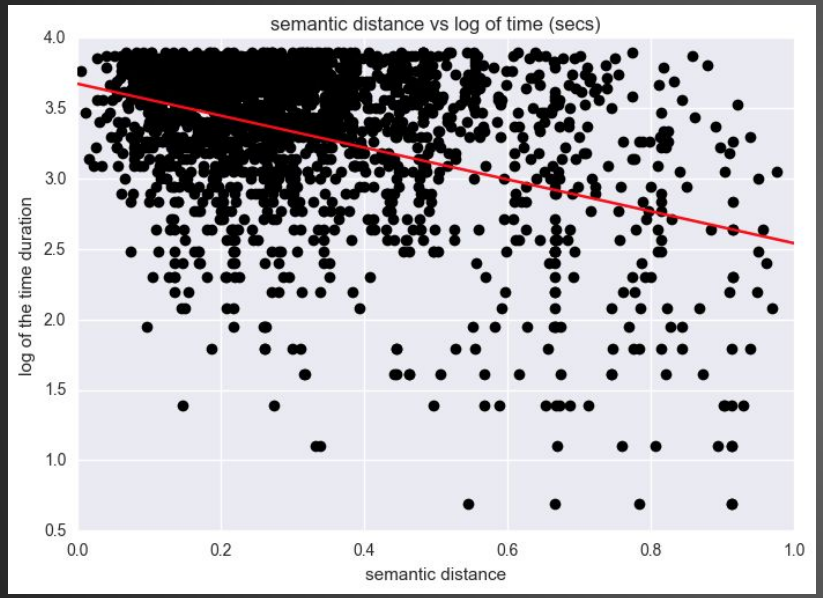


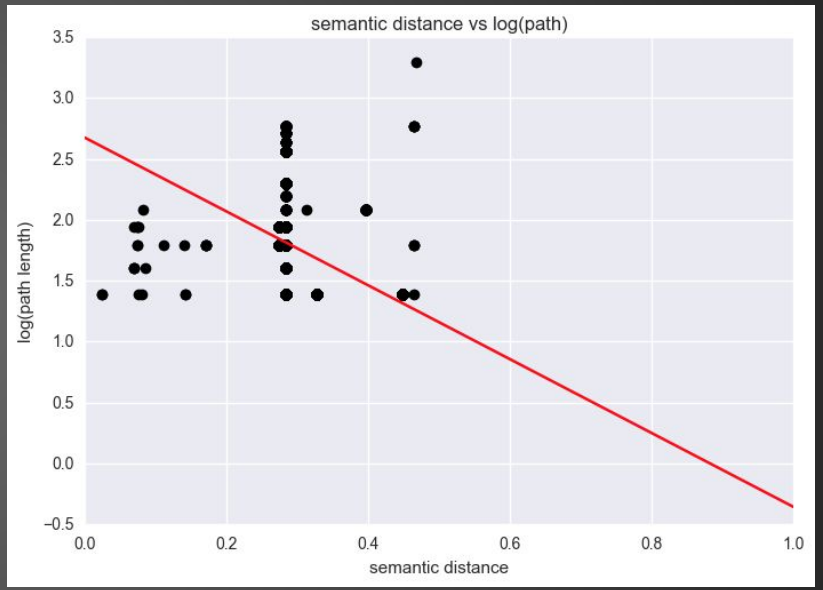
Higher optimal path lengths have weaker semantic connection

Latent Semantic Analysis: Review

- Examples
  - Same article to same article is 1
  - Glasgow to London is .589
  - 14th Century to Ancient Greece is .447
  - Weaknesses
- LSA creates a symmetric matrix, but topics may not be symmetric
  - Example: Minneapolis and Minnesota

Regression Tasks: Results

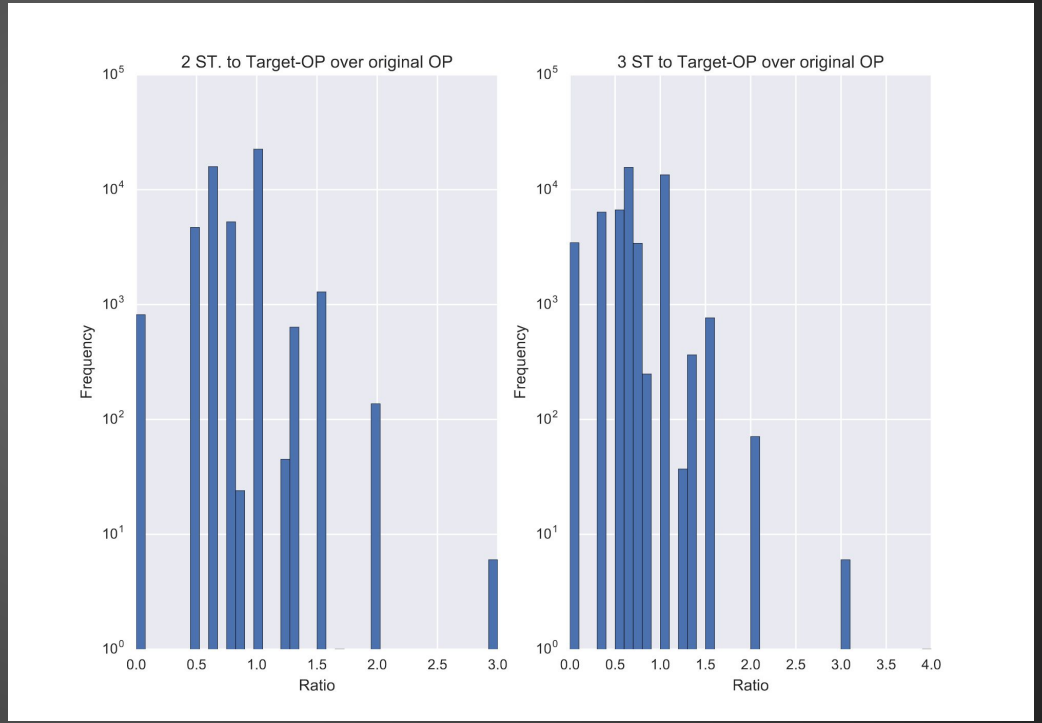




Semantic distance does not correlate to either game duration or path length (R^2 of .18 and .13 respectively)

Summary

- Topic Modeling
  - LDA provides insights into source/target categories which helps us understand relationship between different clusters
- Regression
  - No correlation between semantic path and game duration
- Next steps
  - Do players get better or worse as they navigate the path?



References

- http://snap.stanford.edu/data/wikispeedia.html
- http://infolab.stanford.edu/~west1/pubs/West-Leskovec_WWW-12.pdf
- http://infolab.stanford.edu/~west1/pubs/West-Pineau-Precup_IJCAI-09.pdf
- https://cseweb.ucsd.edu/~jmcauley/cse255/reports/wi15/Shelby_Thomas_Moein_Khazraee.pdf

Thank you! Questions?