# Big Metadata Toolkits:

## Pattern Analysis and Categorization Components

Joseph A Busch, Taxonomy Strategies


Copyright 2015 Taxonomy Strategies LLC. All rights reserved.

# What is big data

- Big data is really the aggregation of metadata
  - E.g., phone call logs, or other automatically generated transaction information.
- To be able to do analytics, and identify emergent patterns, metadata is required.
- The problem is how to efficiently organize and tag content, or generate complete and consistent metadata.

# Indexer inconsistency

- Studies have consistently shown that levels of consistency vary, and that high levels of consistency are rare for:
  - Indexing
  - Choosing keywords
  - Prioritizing index terms
  - Choosing search terms
  - Assessing relevance
  - Choosing hypertext links

**30%**

- Semantic tools and automated processes can help guide users to be more consistent.

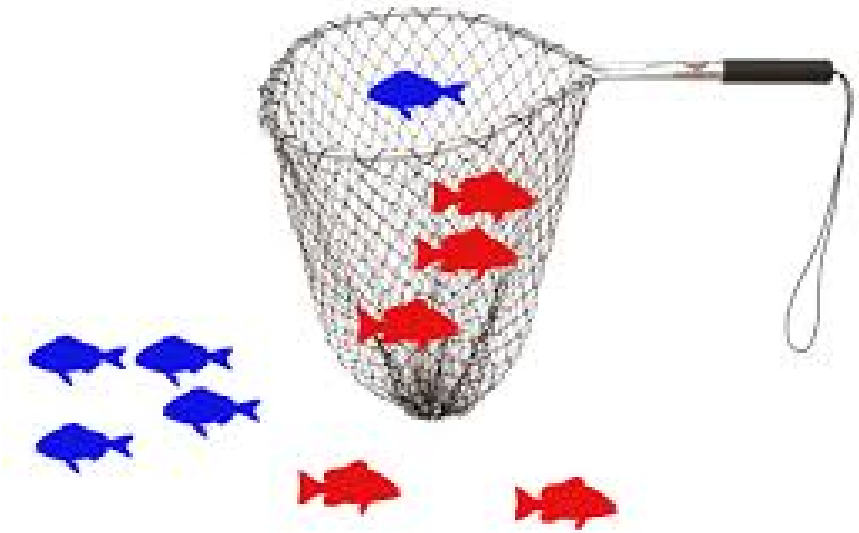
**80%**

---

Markey, K. "Inter-indexer consistency tests: A literature review and report of a test of consistency in indexing visual materials." *Library and Information Science Research*, 6, 155-177, 1984.

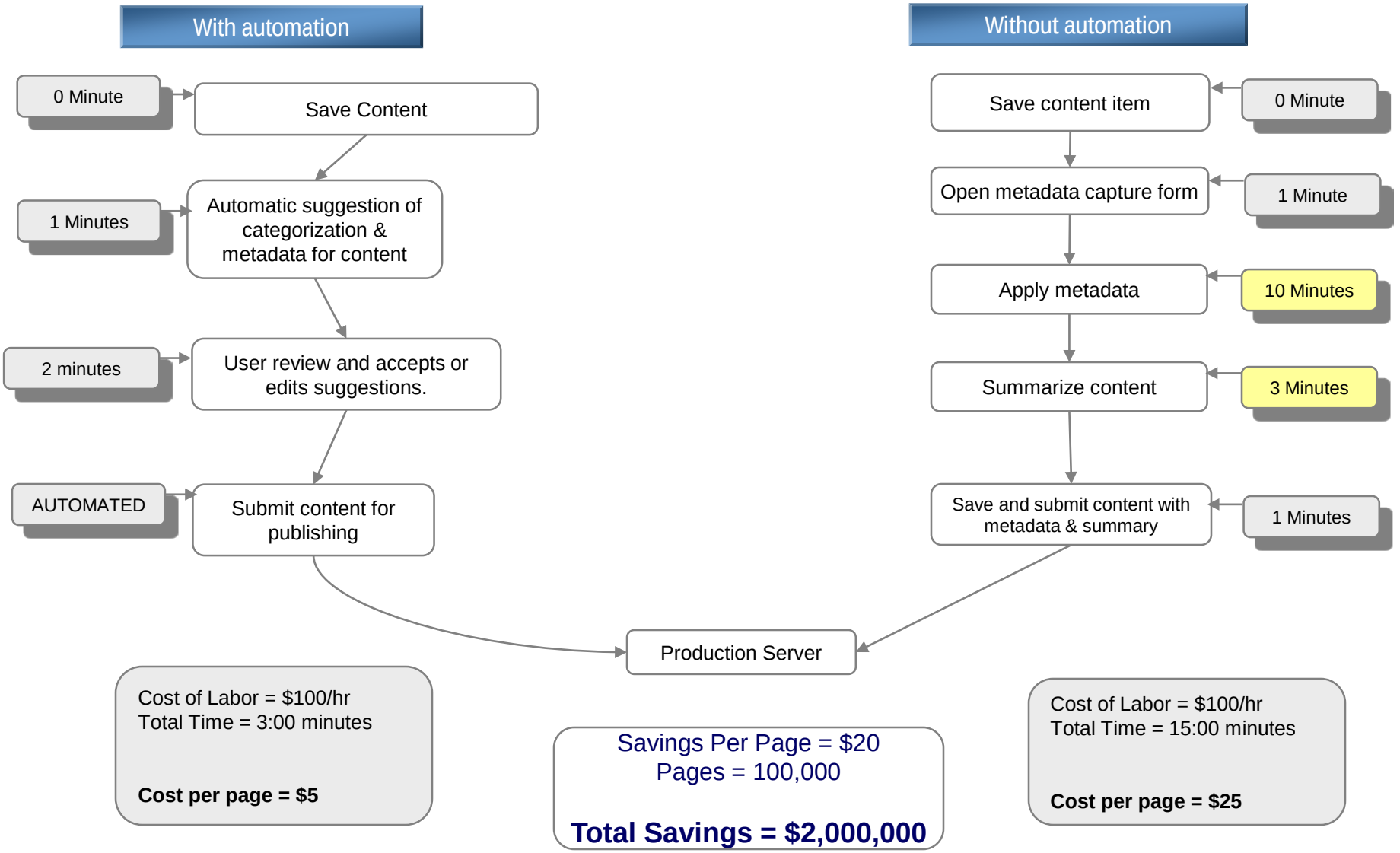
# Precision and Recall

- In categorization there is always a tradeoff between accuracy and completeness.



Red fish are relevant. Blue fish are not relevant. The net is the query.
http://www.lucidatainc.com/2012/10/evaluating-accuracy-of-your-ediscovery-searches-queries-and-software/

# Best case scenario:
# Automated suggestion + SMEs review and improve



# Big metadata toolkit

- Keyword and regular expression matching
- Templates and business rules
- Complex pattern categorizers
- Entity extraction
- Trained categorizers
- Vector space model

# Keyword and regular expression matching

- Category: Automobiles
  - Keywords: car, auto
  - 382 words, 7 matches, term frequency 1.83%

A Cadillac de Ville prototype, presented to Charles E. Wilson upon his retirement as General Motors chief executive in 1953, will be seen in public for the first time in about 50 years, at the Amelia Island Concours d'Élégance in Florida in March.

Mr. Wilson, forced to retire when he accepted President Dwight D. Eisenhower's appointment to become Secretary of Defense, was famously misquoted as saying, "What's good for General Motors is good for the country" during his confirmation hearings. He was nicknamed Engine Charlie to distinguish him from Charles E. Wilson, the chief executive of General Electric, who served in the Truman administration.

The Fleetwood-bodied Coupe de Ville prototype was built for G.M.'s Transportation Unlimited auto show in 1949. The exhibition, which later became known as the Motorama, toured the auto show circuit to great acclaim and heavy attendance. The prototype later given to Wilson was one of four built and the only one known to survive.

The Coupe de Ville, powered by the company's first overhead valve V-8, represented a revolutionary new post-World War II styling statement from G.M. It cost a reported $30,000 – nearly $300,000 today – and took more than two months to build. It is the oldest known Motorama vehicle still extant, according to the concours.

The Caddy's curved, one-piece windshield glass was remarkable for its day. The prototype also boasted fanciful features that may seem quaint by today's automotive standards, but were almost in the realm of fantasy then: A two-way radio/telephone, power windows that included even the vent windows, power seats, chrome wheel arches, a three-piece rear window, a lipstick holder, a perfume atomizer, a rear-seat secretarial kit and leather seats and trim.

The car's current owner, Steve Plunkett of London, Ontario, is the owner of 49 rare and historically significant Cadillacs. He said the Coupe de Ville had recently been completely and correctly restored.

The car had disappeared after Mr. Wilson's death in 1961. It was found in a Connecticut barn in 1978.

"I have many cars, but the Coupe de Ville restoration has been the most exciting," Mr. Plunkett said in an interview. "This will be the first time in 64 years the oldest surviving Motorama Dream Car will be displayed publicly."

It will be judged in the Concept Cadillac class at the concours.

# Improving accuracy through expanded key- words and phrases

| Type | Values | Scope Note |
|---|---|---|
| Category: | Automobiles | a motor vehicle with four wheels; usually propelled by an internal combustion engine |
| Keywords: | automobile, car, auto, motorcar | |
| Hyponyms: | compact, compact car | a small and economical car |
| | convertible | a car that has top that can be folded or removed |
| | coupe | a car with two doors and front seats and a luggage compartment |
| | … | |
| Brands: | Buick, Cadillac, Chevrolet, Chrysler, Ford, General Motors, G.M., Jeep, … | American automobile brands and company names |

# Keyword and regular expression matching with improved accuracy

- Category: Automobiles
  - Keywords: car, auto, coupe, Cadillac, General Motors, G.M.
  - 382 words, 17 matches, term frequency 4.45%

A Cadillac de Ville prototype, presented to Charles E. Wilson upon his retirement as General Motors chief executive in 1953, will be seen in public for the first time in about 50 years, at the Amelia Island Concours d'Élégance in Florida in March.

Mr. Wilson, forced to retire when he accepted President Dwight D. Eisenhower's appointment to become Secretary of Defense, was famously misquoted as saying, "What's good for General Motors is good for the country" during his confirmation hearings. He was nicknamed Engine Charlie to distinguish him from Charles E. Wilson, the chief executive of General Electric, who served in the Truman administration.

The Fleetwood-bodied Coupe de Ville prototype was built for G.M.'s Transportation Unlimited auto show in 1949. The exhibition, which later became known as the Motorama, toured the auto show circuit to great acclaim and heavy attendance. The prototype later given to Wilson was one of four built and the only one known to survive.

The Coupe de Ville, powered by the company's first overhead valve V-8, represented a revolutionary new post-World War II styling statement from G.M. It cost a reported $30,000 – nearly $300,000 today – and took more than two months to build. It is the oldest known Motorama vehicle still extant, according to the concours.

The Caddy's curved, one-piece windshield glass was remarkable for its day. The prototype also boasted fanciful features that may seem quaint by today's automotive standards, but were almost in the realm of fantasy then: A two-way radio/telephone, power windows that included even the vent windows, power seats, chrome wheel arches, a three-piece rear window, a lipstick holder, a perfume atomizer, a rear-seat secretarial kit and leather seats and trim.

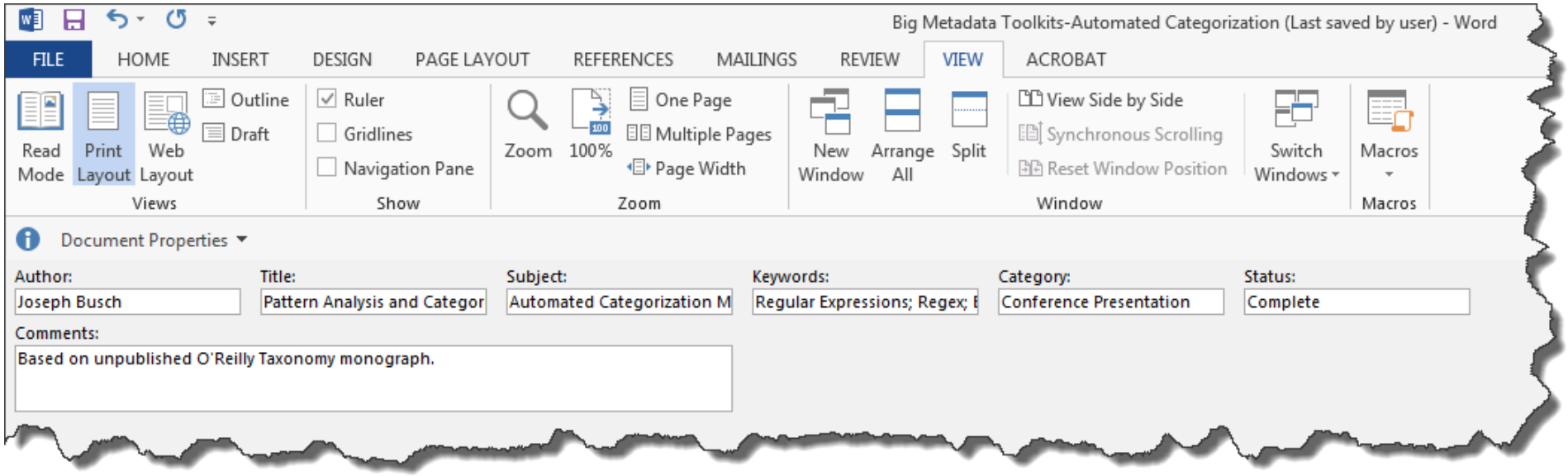
The car's current owner, Steve Plunkett of London, Ontario, is the owner of 49 rare and historically significant Cadillacs. He said the Coupe de Ville had recently been completely and correctly restored.

The car had disappeared after Mr. Wilson's death in 1961. It was found in a Connecticut barn in 1978.

"I have many cars, but the Coupe de Ville restoration has been the most exciting," Mr. Plunkett said in an interview. "This will be the first time in 64 years the oldest surviving Motorama Dream Car will be displayed publicly."

It will be judged in the Concept Cadillac class at the concours.

# Templates and business rules: Manual metadata




- Enter metadata values into template used to add content to the system, e.g., MS Office document properties.

# Metadata capture template: Generic investment products


- Blank metadata capture template, with no values defaulted, all fields used.

# Metadata capture template: Generic investment products


- Metadata capture controlled vocabulary pick list

# Metadata capture template: Contextual defaults


- Products page metadata capture template:
  - Audience defaults to Retail
  - Segment to Direct
  - Portfolio, Strategy, Broad Asset Class, Investment Style, and Risk Level fields are hidden.

# Templates and business rules: Simple rules

- Tag top-level content first
  - Tag landing pages for major sections
  - Lower-level pages inherit tags from top-level pages
- **If** content originated in this department, **then** tag it with pre-defined values.
- **If** the first line of content is centered and in title case, **then** use it to fill-in the Title field.
- Assume that the person who is logged on is the <creator> of the content
  - Inherit the department in which that person works as the content <publisher>

# Complex pattern categorizers: Scoring/weighting

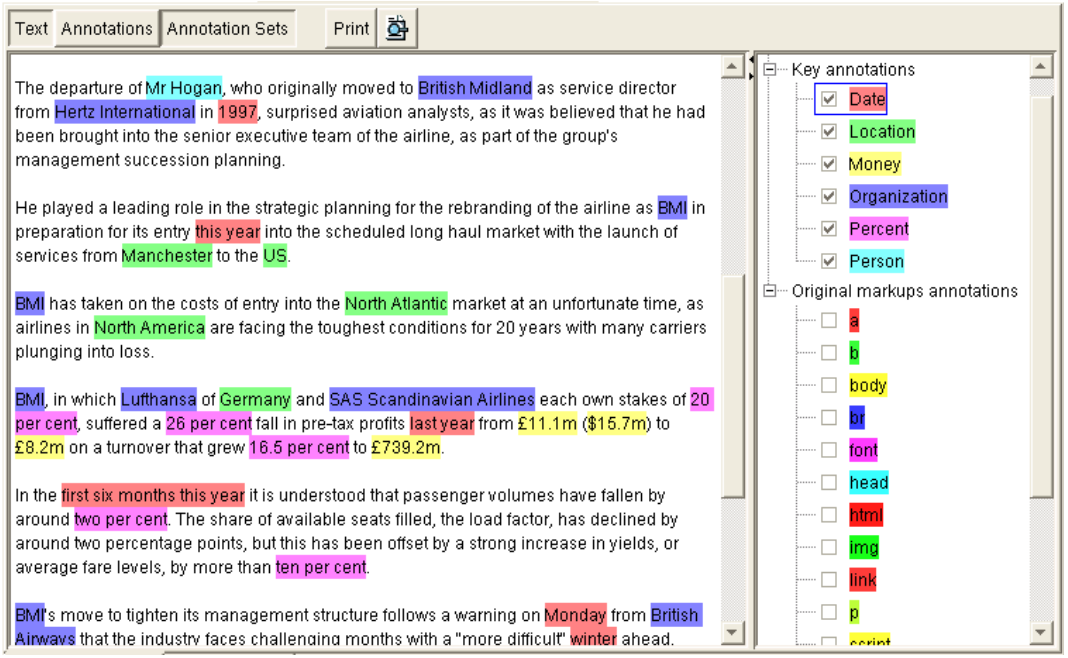
- Count number of matches for each category within a content item, then assign item to the most frequently occurring category.
- Weight matches in the title more heavily than matches in the body.
- Weight phrase matches more heavily than single word matches.
- Use negative matches (exclusions) to reduce false positives.
  - E.g., Match “financial institutions” category to “bank,” but not “river bank” or “Georges Bank”
- Use surrounding context (clues) to boost or reduce score of a match.

# Complex pattern categorizers: More complex rules

- Each category needs its own rules to describe the things that match the category.
  - Start very simple, frequently just the category name and a synonym or two.
  - Iteratively analyze ever-increasing amounts of content.
  - Check for false negatives and false positives.

# Entity extraction

- Processing text to identify the mentions of entities such as people, places, organizations, and products, as well as addresses, dates, currency amounts, job titles, and topics.
- Like other text matches, occurrences of named entities can be weighted and scored with a threshold set to assign them as meaningful categories for a content item.



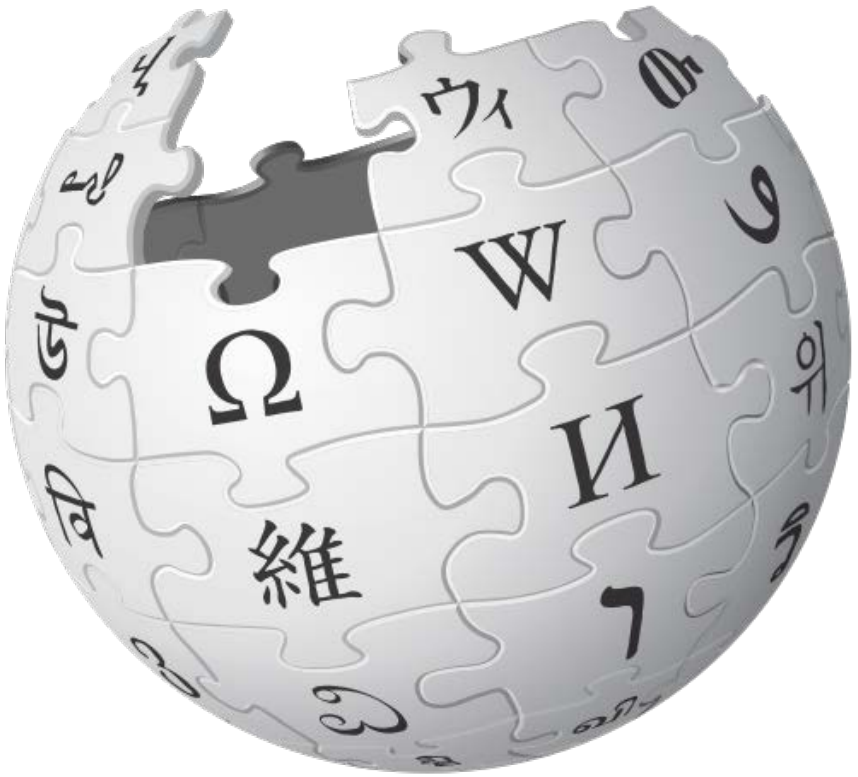

ANNIE - Open Source Information Extraction (http://www.aktors.org/technologies/annie/)

# Entity extraction enhancements

- Use POS tagging of the content, and clues such as patterns in the neighborhood of the text to identify noun phrases that are more likely to be named entities.
- Use large authority files—lists of the names of people, places, organizations, products, and so forth which also include variations on those names, providing patterns to be recognized that are not obvious variations.
- Make use of heuristics
  - A number followed by a proper name that ends in St., Dr., or Blvd. is probably a street address.
  - A noun phrase ending in a gerund, such as Minnesota Mining and Manufacturing, is probably an organization name.

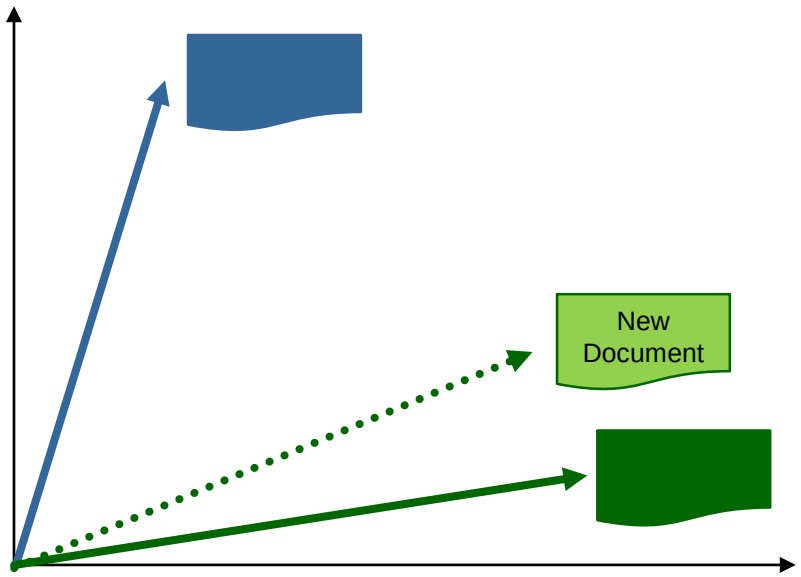
# Trained categorizers

- A training set is a collection of pre-categorized items.
  - E.g., The articles in Wikipedia
- The training set can be analyzed to identify the words that occur and co-occur in each category.
- A new document can have its word information analyzed, and then tagged with the category that is best matched.



# Vector Space Model

- Each document is represented by a term vector, a column of numbers generated by analyzing the document content
  - E.g., tf-idf (term frequency-inverse document frequency) number of times each word appears in the document, offset by the frequency of each word in a collection of documents
- To categorize a new document, calculate its vector and identify which one is closest.



# Summary

- To be able to do analytics, and identify emergent patterns, metadata is required.
- We have discussed a variety of methods based on are built on Regex and authority files to efficiently organize and tag content, or generate complete and consistent metadata.

Joseph A Busch, Principal
jbusch@taxonomystrategies.com
twitter.com/joebusch
415-377-7912

# QUESTIONS?

# Description

- The exploding volume, complexity and velocity of structured and unstructured data, especially from all sorts of interactions with it, presents challenges and opportunities to derive valuable insights. Among the challenges of managing massive data sets are gathering, validating, preserving, analyzing and maintaining linkages from those analyses to the source data set. Identifying patterns in data sets using information retrieval methods and writing out the results, as metadata is a well established information management process that should be adopted by organizations working with today's big data sets. This presentation provides an overview of pattern analysis and categorization methods.