Website Categorisation

Joseph Atkins-Turkish Jake Hodges James Patrick-Evans

May 5, 2015

The Idea

The aim is to take a website and automatically assign it to a descriptive category label.

- Original idea to sort bookmarks into folders
- Categorisation of websites
  - Organising bookmarks
  - Automated SEO
  - Content filtering
- Supervised multi-class classification problem
  - Unsupervised element

Approach

1. Data Collection
   - Compile corpus[1] of websites with pre-existing class labels
2. Data Sanitisation
   - Process and sanitise raw HTML
3. Feature Extraction
   - Find key features in corpus (LDA or tf-idf)
4. Classification
   - Use generated features to categorise websites

[1] a collection of documents

Approach

1. Data Collection
   - Compile corpus[1] of websites with pre-existing class labels
2. Data Sanitisation
   - Process and sanitise raw HTML
3. Feature Extraction
   - Find key features in corpus (LDA or tf-idf)
4. Classification
   - Use generated features to categorise websites

[1] a collection of documents

Approach

1. Data Collection
   - Compile corpus[1] of websites with pre-existing class labels
2. Data Sanitisation
   - Process and sanitise raw HTML
3. Feature Extraction
   - Find key features in corpus (LDA or tf-idf)
4. Classification
   - Use generated features to categorise websites

[1] a collection of documents

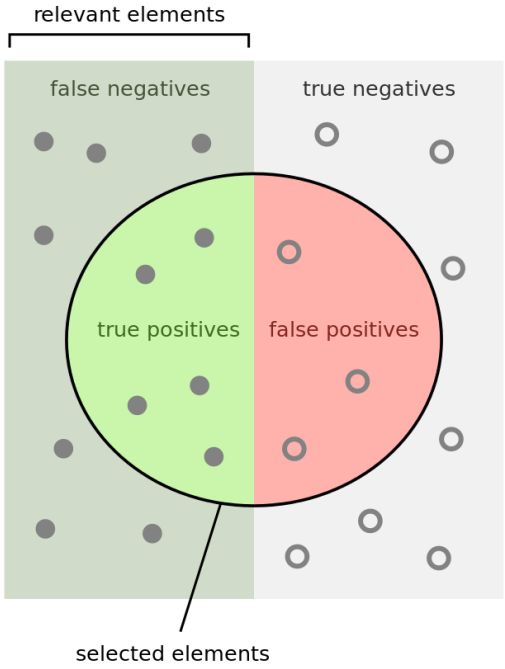
Approach

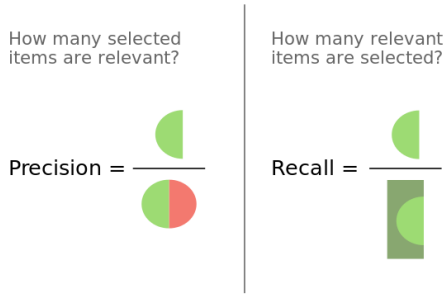
1. Data Collection
   - Compile corpus[1] of websites with pre-existing class labels
2. Data Sanitisation
   - Process and sanitise raw HTML
3. Feature Extraction
   - Find key features in corpus (LDA or tf-idf)
4. Classification
   - Use generated features to categorise websites

[1] a collection of documents

F1 Metric

- Need an objective metric to evaluate results of classification
- Removes biases due to non-uniform class distribution
- A weighted average of the *precision* and *recall*
- Also average of F1 score of each class, weighted by class proportions





Compiling a corpus (1)

Data Collection

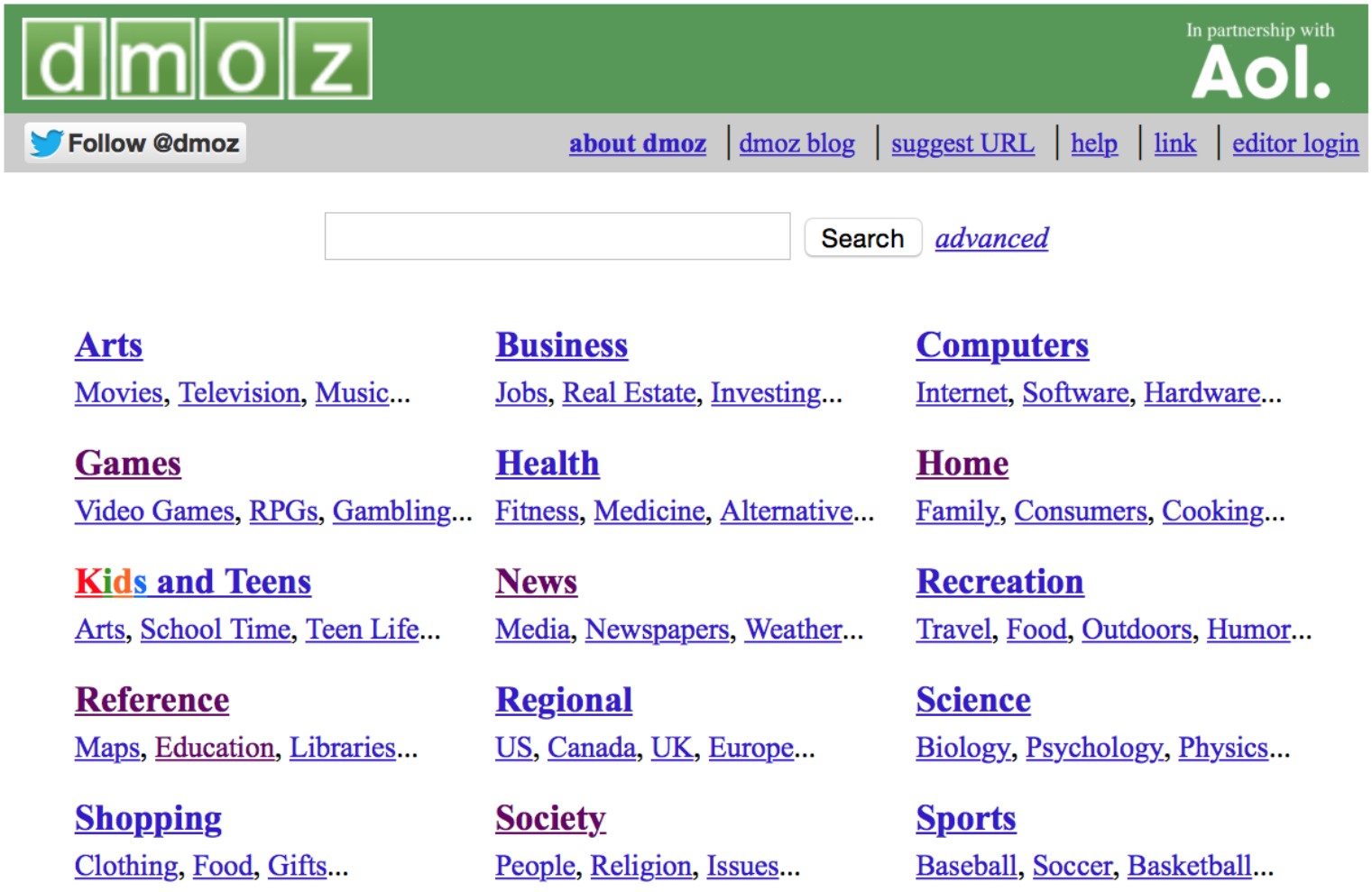
Require corpus of websites arranged into hierarchical categories

- 20 Newsgroups
- dmoz.org
- Reuters-21578

Compiling a corpus (2)

Data Collection

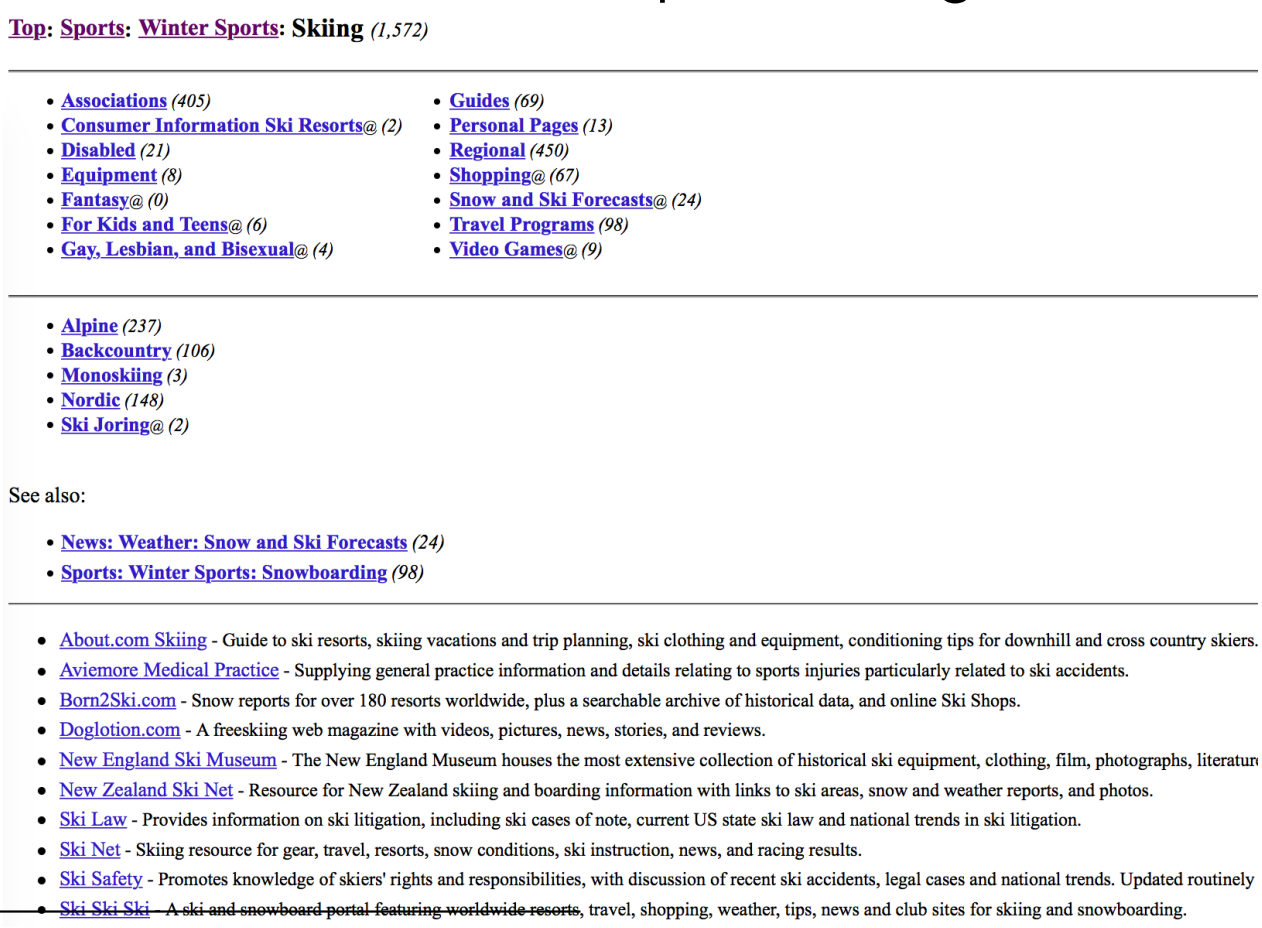
dmoz is a large, hierarchical web directory curated by humans.



Compiling a corpus (3)

Data Collection

We built a scraper using Scrapy[2] and used it to scrape 2000 websites from 15 top-level categories.



Top: **Sports**: **Winter Sports**: **Skiing** *(1,572)*

- **Associations** *(405)*
- **Consumer Information Ski Resorts**@ *(2)*
- **Disabled** *(21)*
- **Equipment** *(8)*
- **Fantasy**@ *(0)*
- **For Kids and Teens**@ *(6)*
- **Gay, Lesbian, and Bisexual**@ *(4)*
- **Guides** *(69)*
- **Personal Pages** *(13)*
- **Regional** *(450)*
- **Shopping**@ *(67)*
- **Snow and Ski Forecasts**@ *(24)*
- **Travel Programs** *(98)*
- **Video Games**@ *(9)*

- **Alpine** *(237)*
- **Backcountry** *(106)*
- **Monoskiing** *(3)*
- **Nordic** *(148)*
- **Ski Joring**@ *(2)*

See also:

- **News: Weather: Snow and Ski Forecasts** *(24)*
- **Sports: Winter Sports: Snowboarding** *(98)*

- About.com Skiing - Guide to ski resorts, skiing vacations and trip planning, ski clothing and equipment, conditioning tips for downhill and cross country skiers.
- Aviemore Medical Practice - Supplying general practice information and details relating to sports injuries particularly related to ski accidents.
- Born2Ski.com - Snow reports for over 180 resorts worldwide, plus a searchable archive of historical data, and online Ski Shops.
- Doglotion.com - A freeskiing web magazine with videos, pictures, news, stories, and reviews.
- New England Ski Museum - The New England Museum houses the most extensive collection of historical ski equipment, clothing, film, photographs, literatur
- New Zealand Ski Net - Resource for New Zealand skiing and boarding information with links to ski areas, snow and weather reports, and photos.
- Ski Law - Provides information on ski litigation, including ski cases of note, current US state ski law and national trends in ski litigation.
- Ski Net - Skiing resource for gear, travel, resorts, snow conditions, ski instruction, news, and racing results.
- Ski Safety - Promotes knowledge of skiers' rights and responsibilities, with discussion of recent ski accidents, legal cases and national trends. Updated routinely
- Ski Ski Ski - ~~A ski and snowboard portal featuring worldwide resorts~~, travel, shopping, weather, tips, news and club sites for skiing and snowboarding.

Storing & Querying Documents

Data Collection

- MongoDB, JSON document store
- Apache Mahout
  - LDA
  - Hadoop
  - Scala + Spark
- Hadoop-style Map Reduce
- Process terabytes of data
- REST API for querying corpus


```
intopics=[] alltopics=[]
hierarchy='^sports.foot(.*)$'
projection=[]
```

[1] https://mahout.apache.org/users/clustering/latent-dirichlet-allocation.html

Website Content & Metadata

Data Collection

We take the body text...

And the metadata tags

```
<html xmlns="http://www.w3.org/1999/xhtml"
 xml:lang="en" lang="en">
  <head>
    <title>SkiNet.com: Home of SKI
    Magazine, Skiing Magazine, Warren
    Miller and Nastar</title>
    <meta http-equiv="Content-Type"
    content="text/html; charset=utf-8" />
    <meta name="keywords" content="ski
    magazine, skiing magazine, skinet,
    warren miller, nastar, ski resorts,
    ski gear, skiing videos, skiing photos
    " />
    <meta name="description" content="Ski
    resorts, ski gear reviews, lift
    ticket deals, skiing videos and
    photos and more at SkiNet, home of
    SKI Magazine, Skiing Magazine, Warren
    Miller and Nastar." />
```

Website Content & Metadata

Data Collection

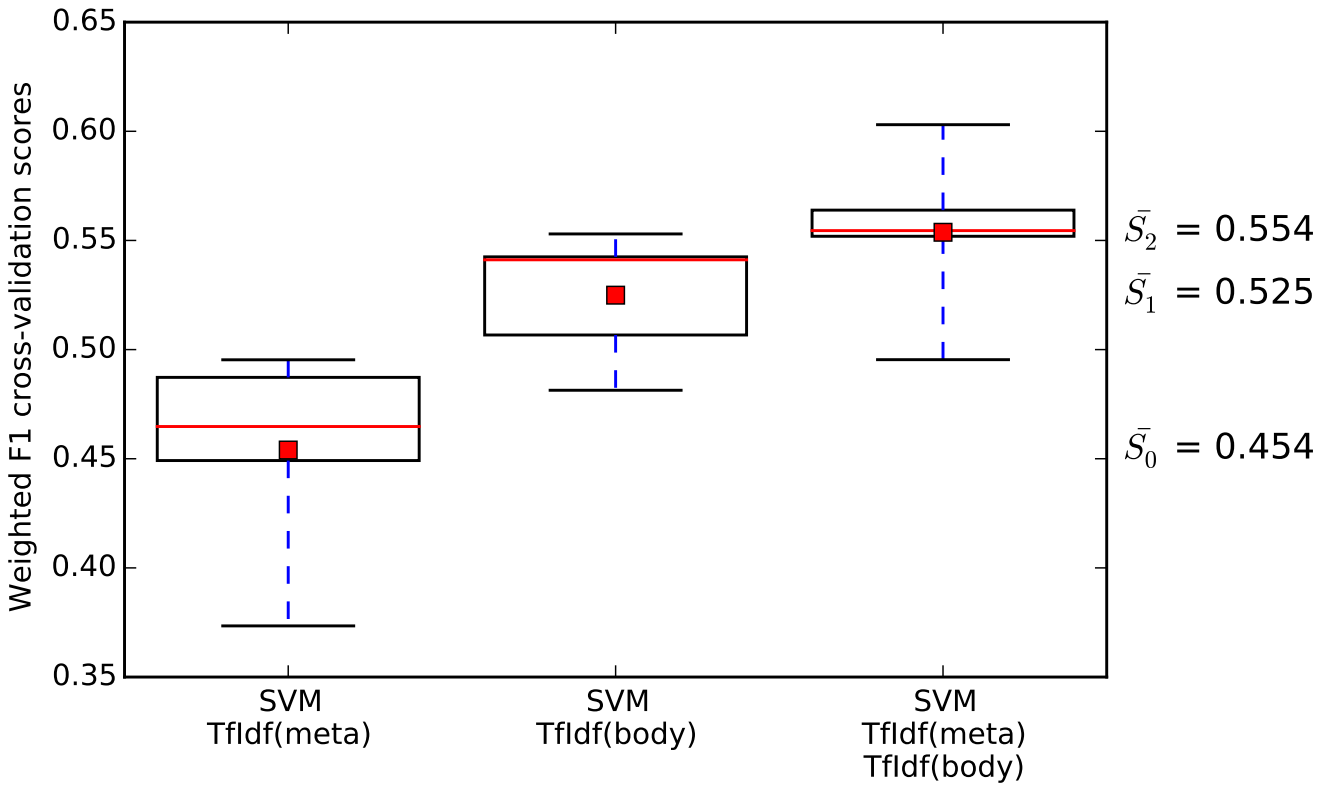
We take the body text...

And the metadata tags

```
<html xmlns="http://www.w3.org/1999/xhtml"
 xml:lang="en" lang="en">
  <head>
    <title>SkiNet.com: Home of SKI
    Magazine, Skiing Magazine, Warren
    Miller and Nastar</title>
    <meta http-equiv="Content-Type"
    content="text/html; charset=utf-8" />
    <meta name="keywords" content="ski
    magazine, skiing magazine, skinet,
    warren miller, nastar, ski resorts,
    ski gear, skiing videos, skiing photos
    " />
    <meta name="description" content="Ski
    resorts, ski gear reviews, lift
    ticket deals, skiing videos and
    photos and more at SkiNet, home of
    SKI Magazine, Skiing Magazine, Warren
    Miller and Nastar." />
```

Metadata vs. Body Text

Data Collection



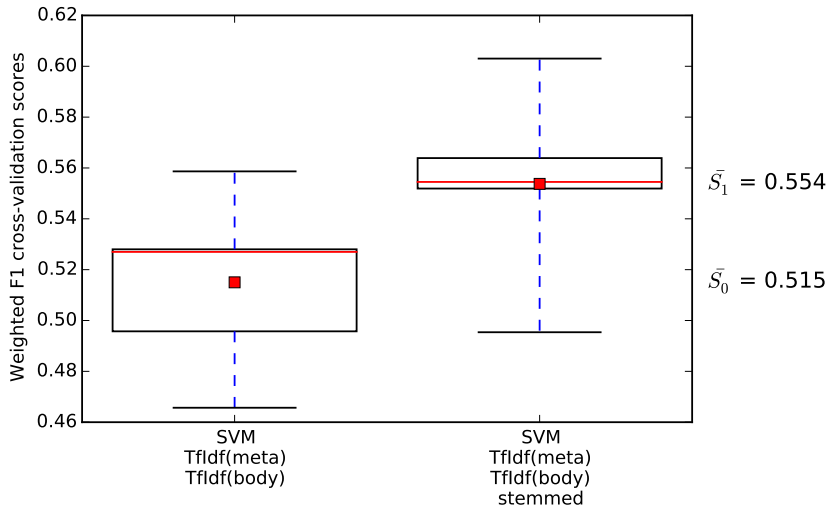
- Body text beats metadata by 0.07 (p-value: 0.028)
- Combining both is an improvement, but only of 0.029 (p-value: 0.22)

Cleaning and Preprocessing

Many words have multiple forms, e.g. plurals. Is it worth counting them as identical features?

- site (2), sites (1), ski (8), skiers (1), skiing (4), skis (1), slog (1) ...
- Use stemming to remove suffixes
  - Ski, Skiing, Skis $\rightarrow$ Ski
  - Skiers $\rightarrow$ Skier
- site (3), ski (13), skier (1), slog (1), ...

F1-score improvement is marginal (0.009), but the algorithm runs 28% quicker.

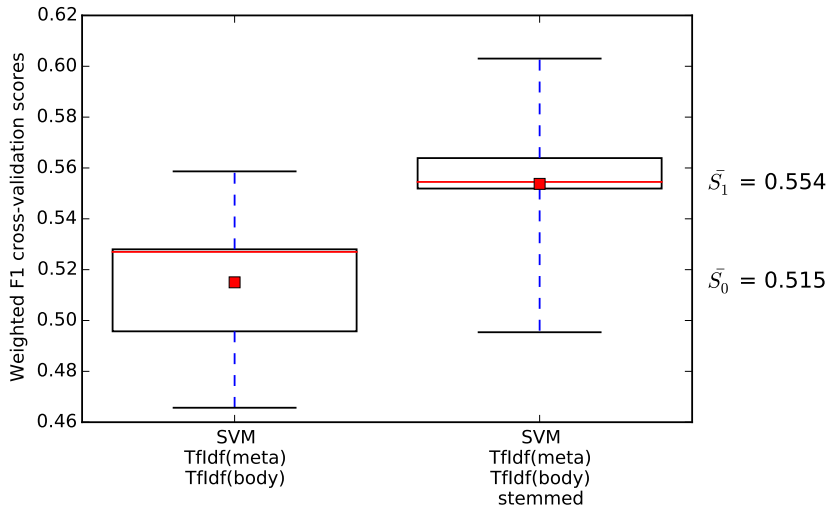


Cleaning and Preprocessing

Many words have multiple forms, e.g. plurals. Is it worth counting them as identical features?

- site (2), sites (1), ski (8), skiers (1), skiing (4), skis (1), slog (1) ...
- Use stemming to remove suffixes
  - Ski, Skiing, Skis $\rightarrow$ Ski
  - Skiers $\rightarrow$ Skier
- site (3), ski (13), skier (1), slog (1), ...

F1-score improvement is marginal (0.009), but the algorithm runs 28% quicker.

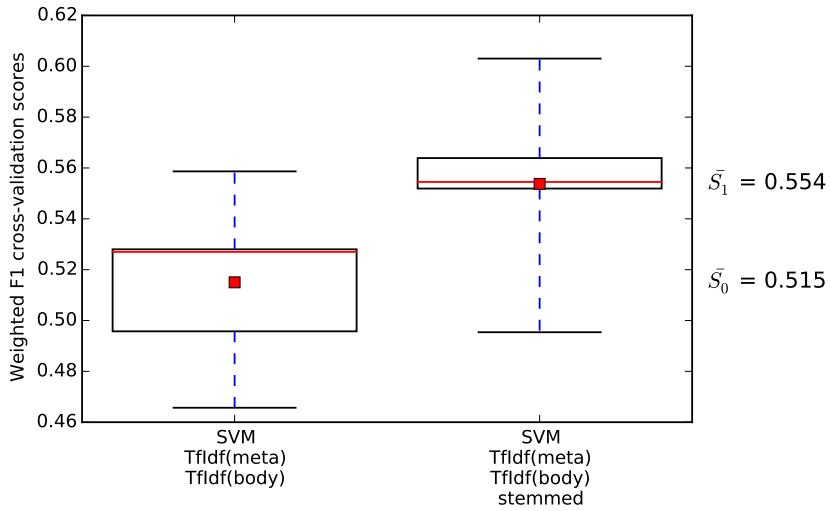


Cleaning and Preprocessing

Many words have multiple forms, e.g. plurals. Is it worth counting them as identical features?

- site (2), sites (1), ski (8), skiers (1), skiing (4), skis (1), slog (1) ...
- Use stemming to remove suffixes
  - Ski, Skiing, Skis $\rightarrow$ Ski
  - Skiers $\rightarrow$ Skier
- site (3), ski (13), skier (1), slog (1), ...

F1-score improvement is marginal (0.009), but the algorithm runs 28% quicker.



Feature Extraction Methods

Feature Extraction

There are many ways to extract useful numerical features from text documents:

- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis (LSA), (and pLSA, LSI, . . .)
- String Kernels
- Term-Frequency Inverse-Document-Frequency (tf-idf)
- Word Counts

Feature Extraction Methods

Feature Extraction

There are many ways to extract useful numerical features from text documents:

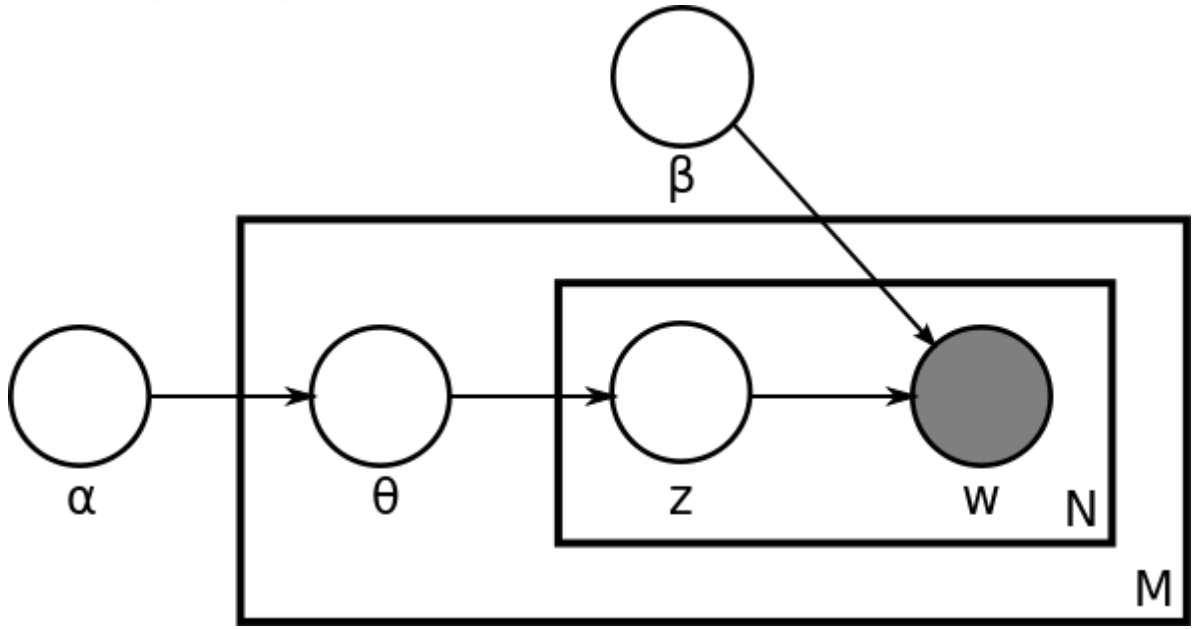
- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis (LSA), (and pLSA, LSI, . . .)
- String Kernels
- Term-Frequency Inverse-Document-Frequency (tf-idf)
- Word Counts

Topic Modelling

Feature Extraction

Latent Dirichlet Allocation (LDA)

- Bag of words model
- Per document topic distributions $\sim Dir(\alpha)$
- Per topic word distribution $\sim Dir(\beta)$
- $Doc\{Topic\{W\}\}$



Term-frequency inverse-document-frequency (tf-idf)

Feature Extraction

$$tfidf(t, d) = \frac{\text{Term Frequency}}{\text{Document Frequency}}$$

$$= \frac{0.5 \times f(t, d)}{max\{f(w, d) : w \in d\}} \times \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

- Importance of word with relation to corpus
- Word count only or just *TermFrequency*, {"*mountain*" : 9}
- Stop words

Classifiers

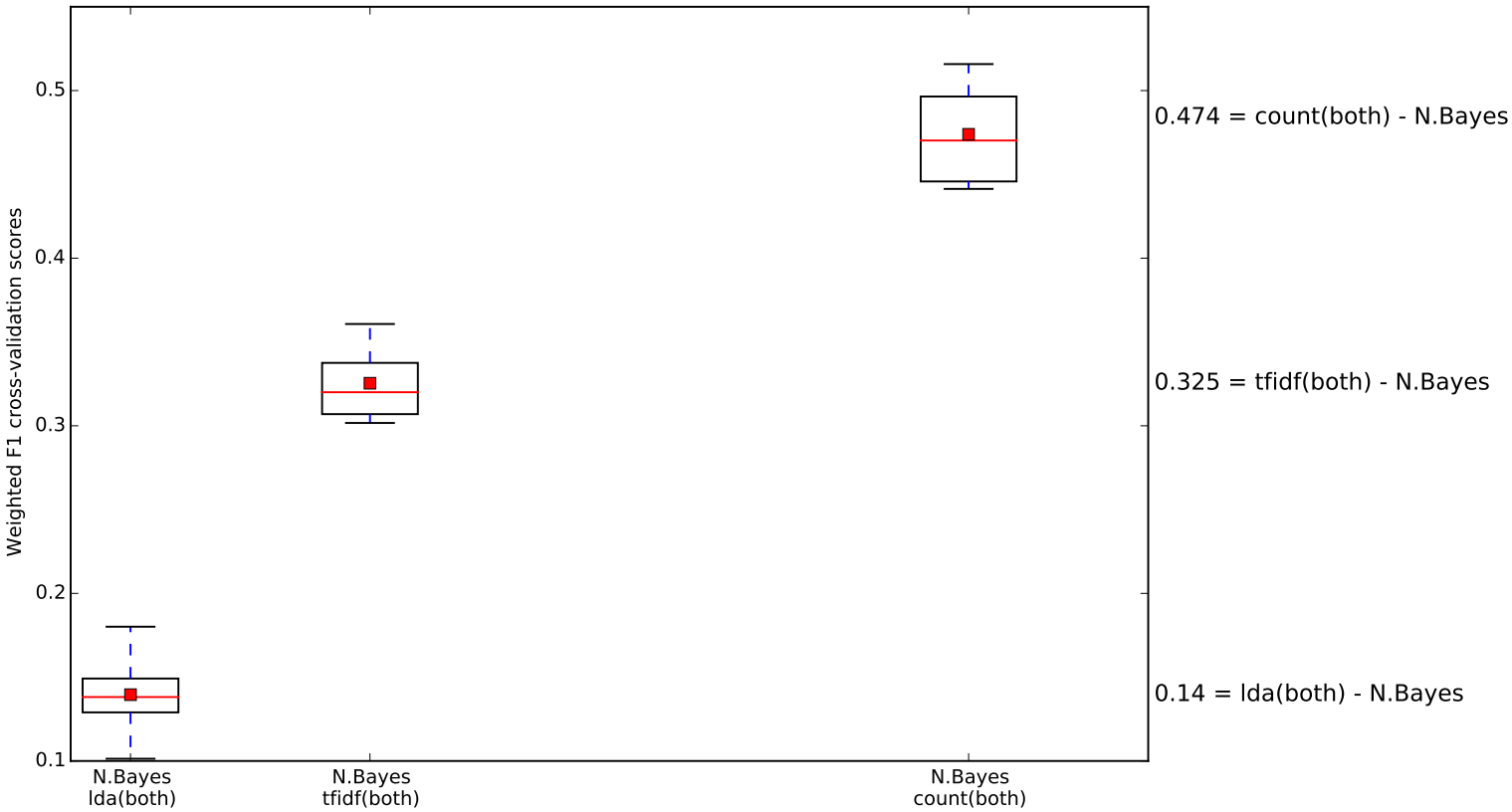
Comparing Classifiers

- Naive Bayes
  - Models classes as Gaussian distributions and classifies based on posterior probability.
- Randomised Trees
  - Ensemble method that fits extremely randomised trees on subsets of the data and averages the results.
- Support Vector Machine Classification
  - Calculates optimal class boundaries by maximising margins between classes.
- Stochastic Gradient Descent optimisation
  - Can be trained 'online' with new data as it comes in
  - Can be used to approximate SVMs, or algorithms which give actual probability estimates.

Results

Classifier comparison

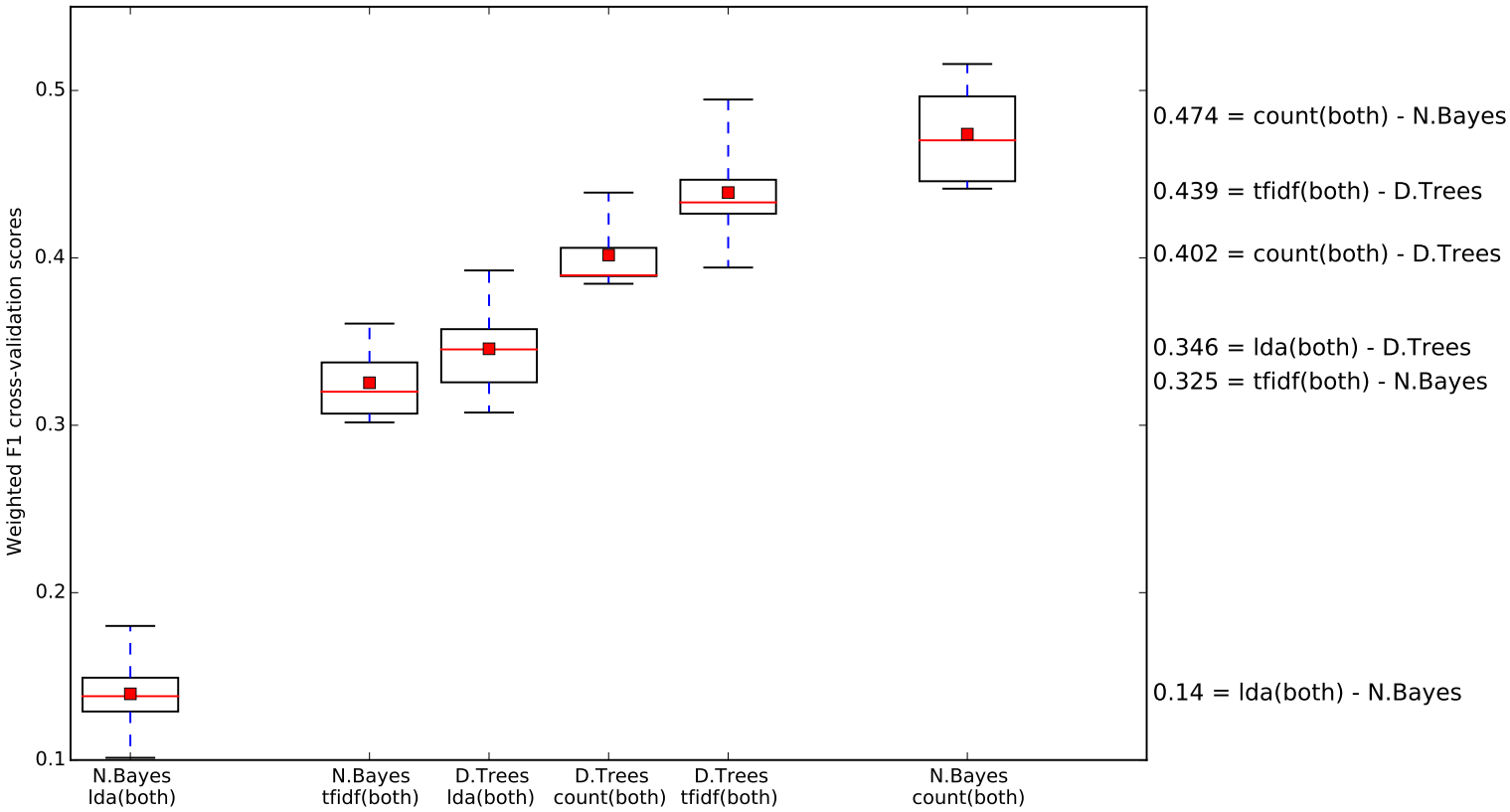
Different classifiers work best with different representations



Results

Classifier comparison

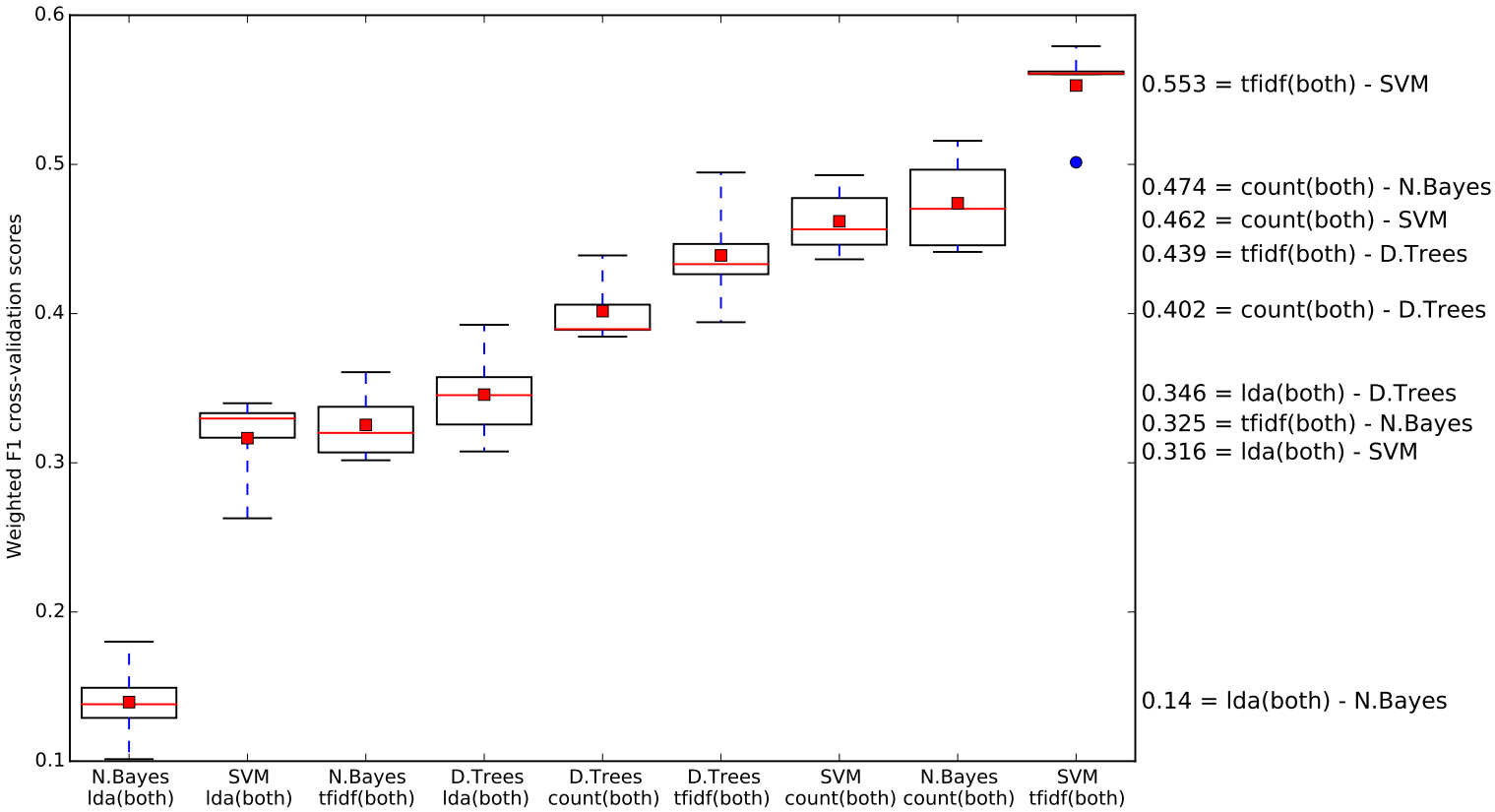
Different classifiers work best with different representations



Results

Classifier comparison

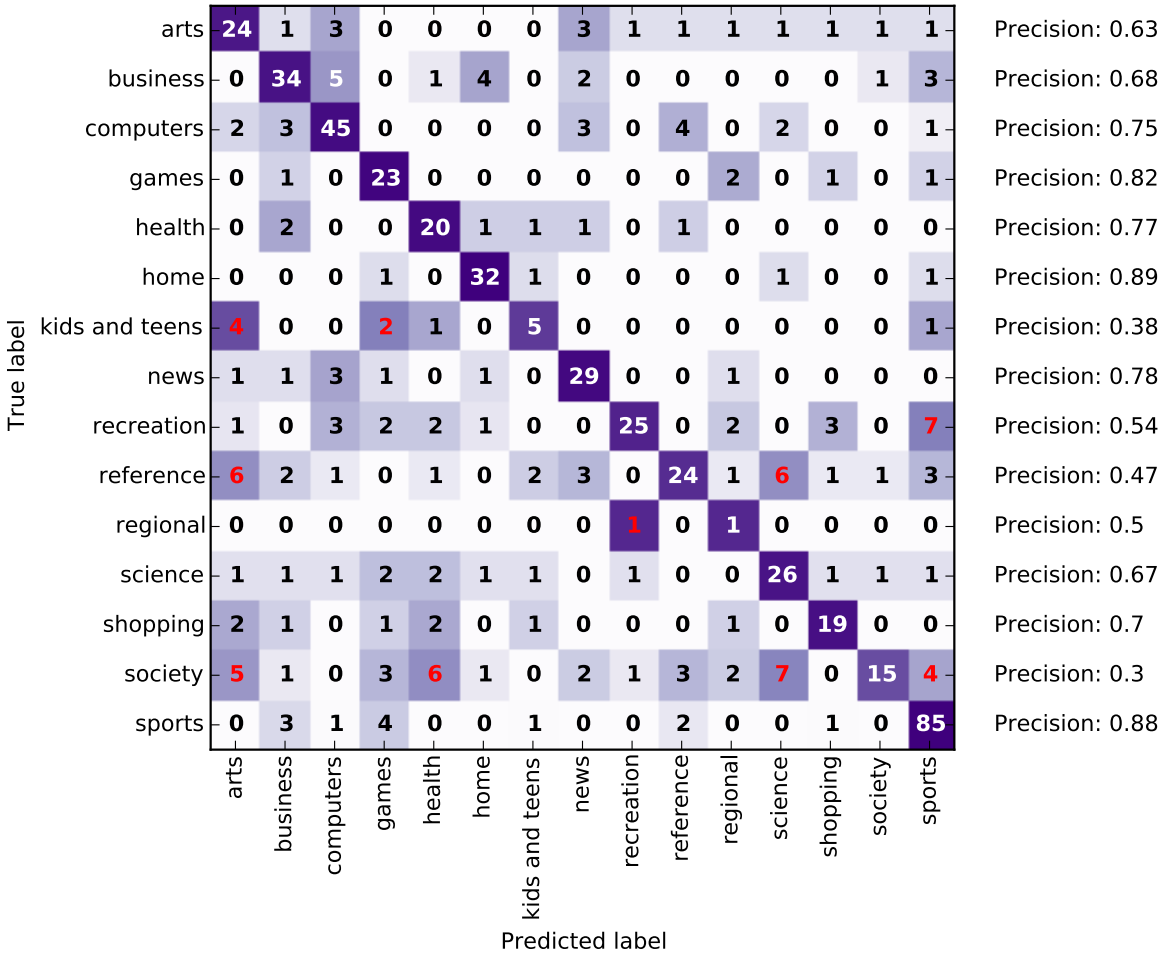
Different classifiers work best with different representations



Confusion Matrices

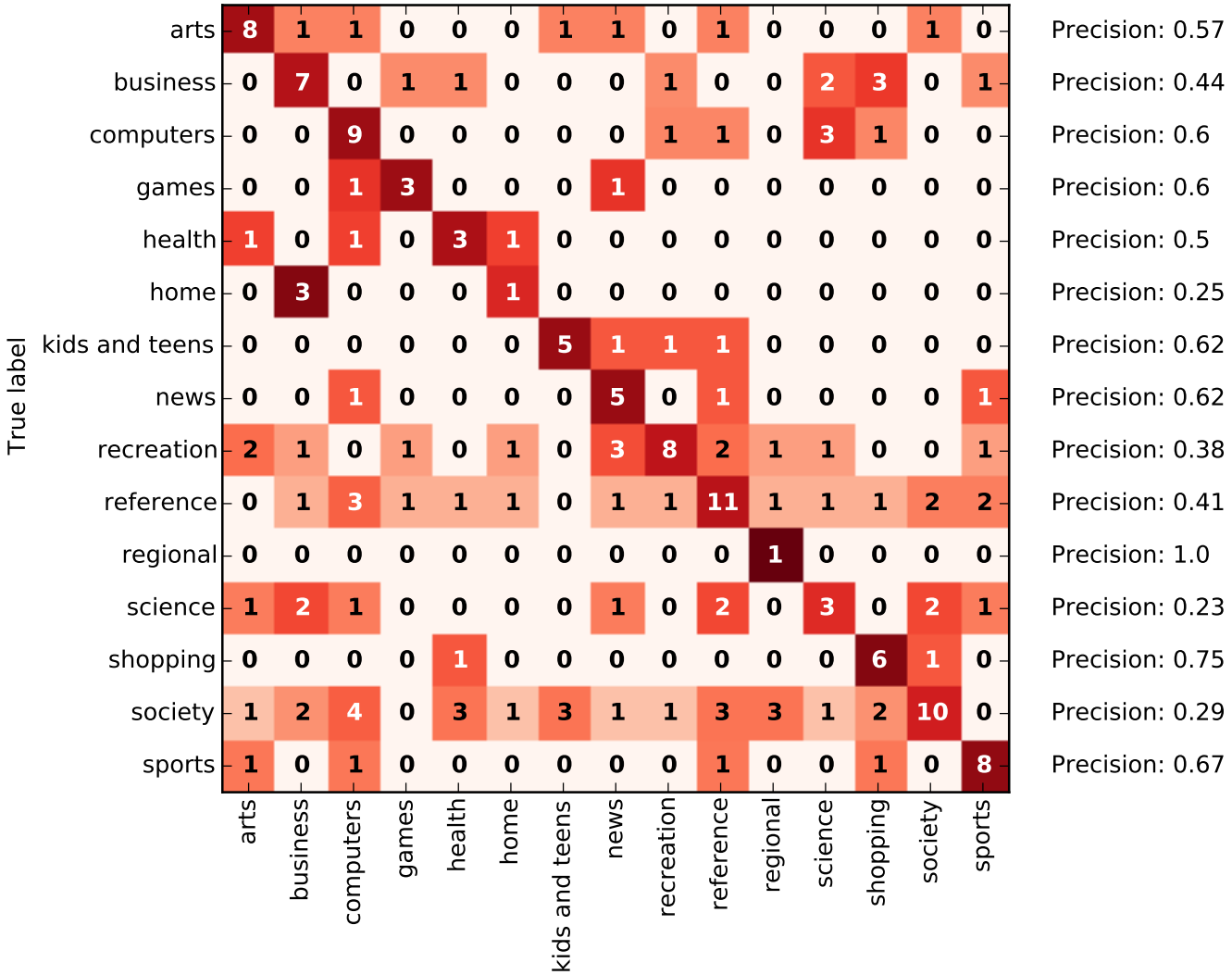
Which things got mistaken for what?

- Confusion matrices let you see where misclassifications happen
- This one is for all data, tf-idf and SVM Classification
- But you can also look at the confusion of *second* guesses...



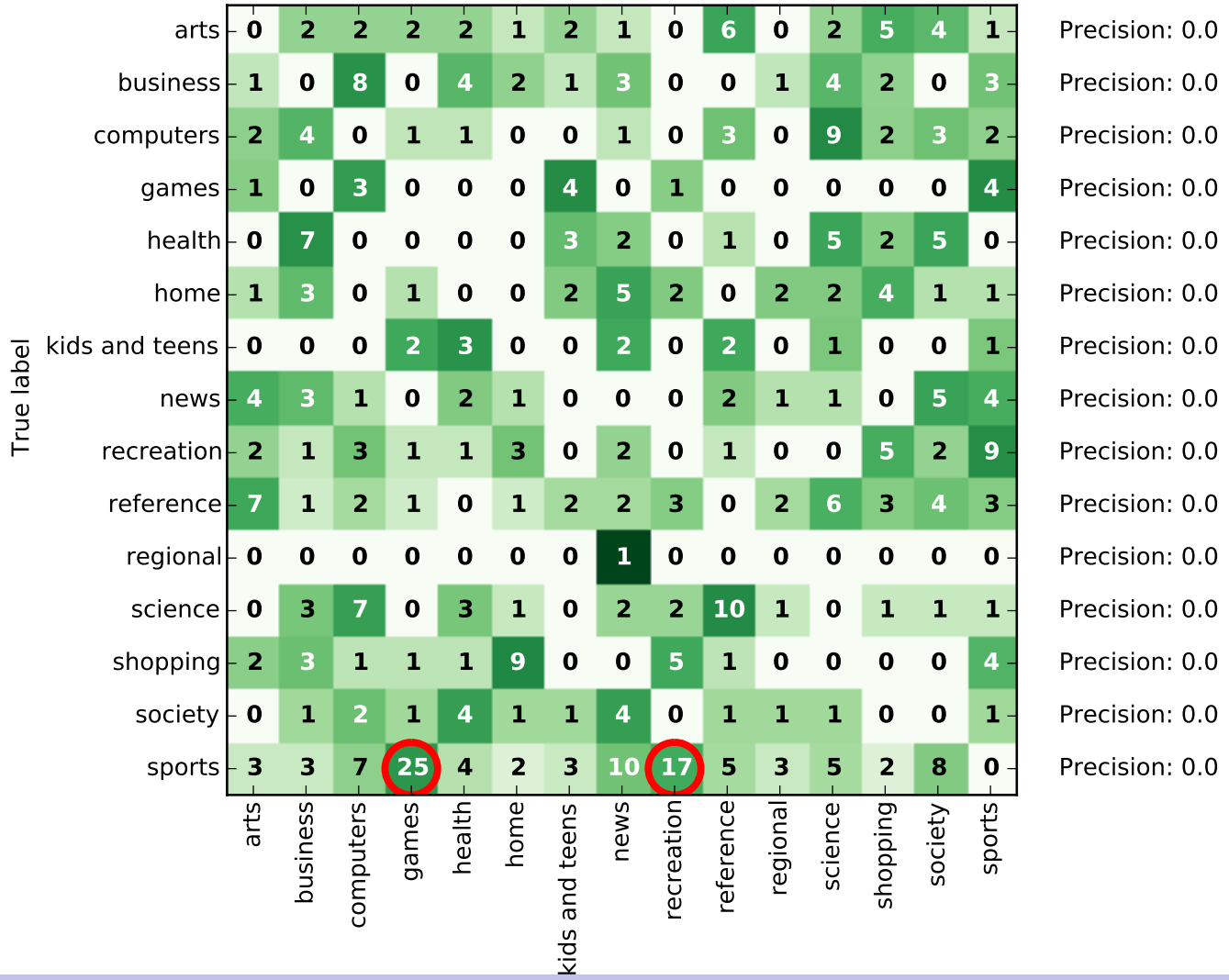
Confusion Matrices

Second guesses when first guesses were wrong

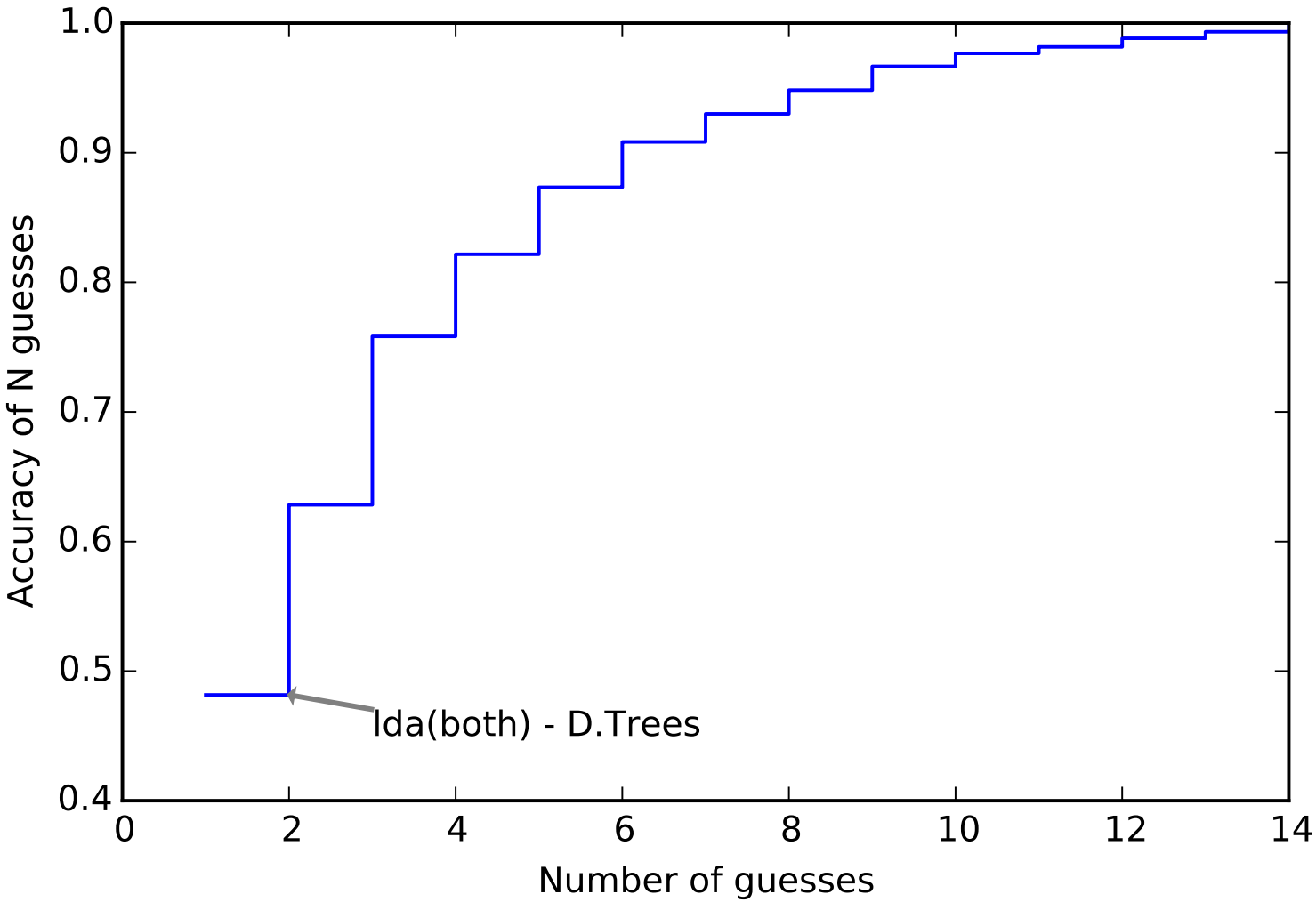


Confusion Matrices

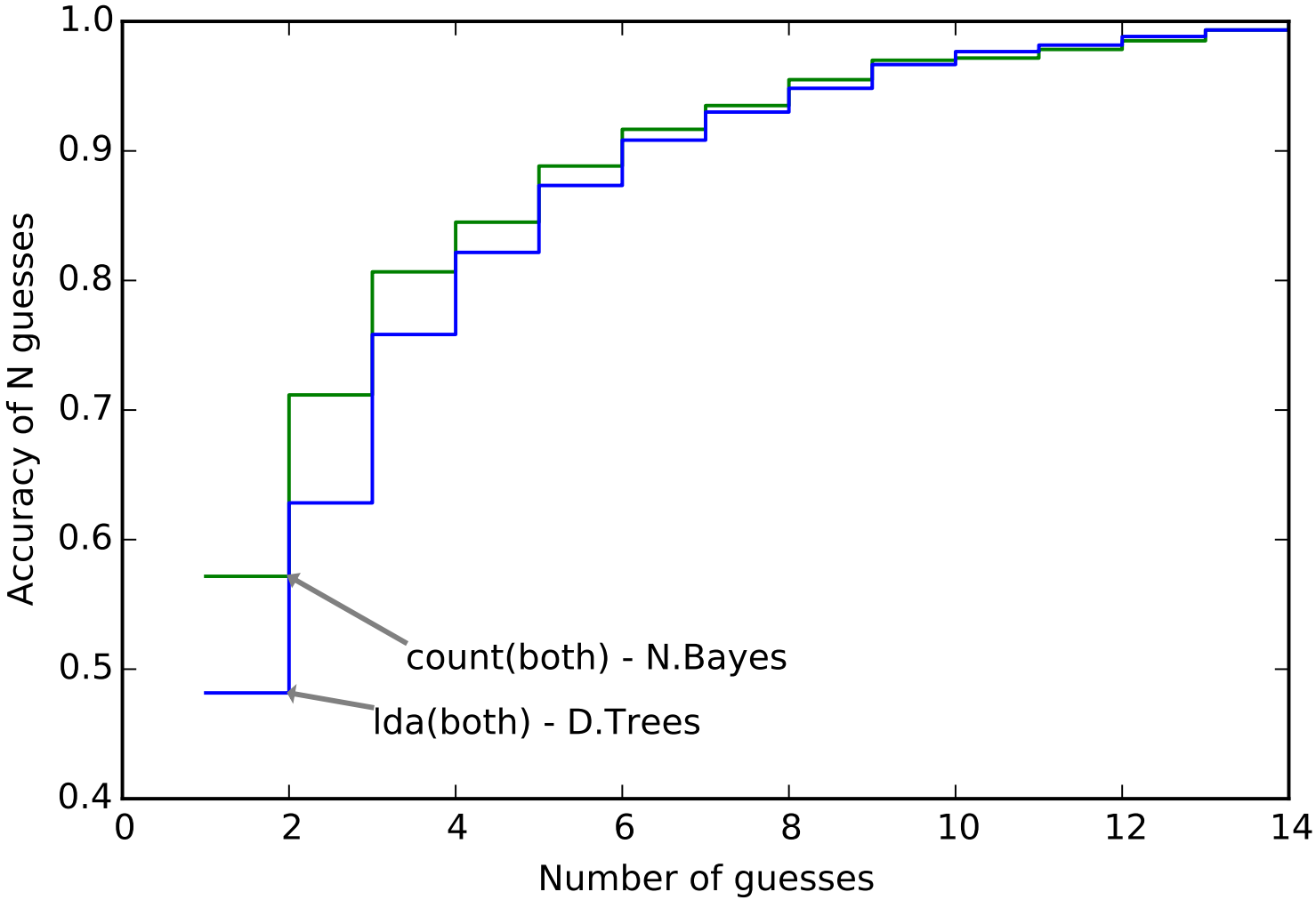
Second guesses when first guesses were right



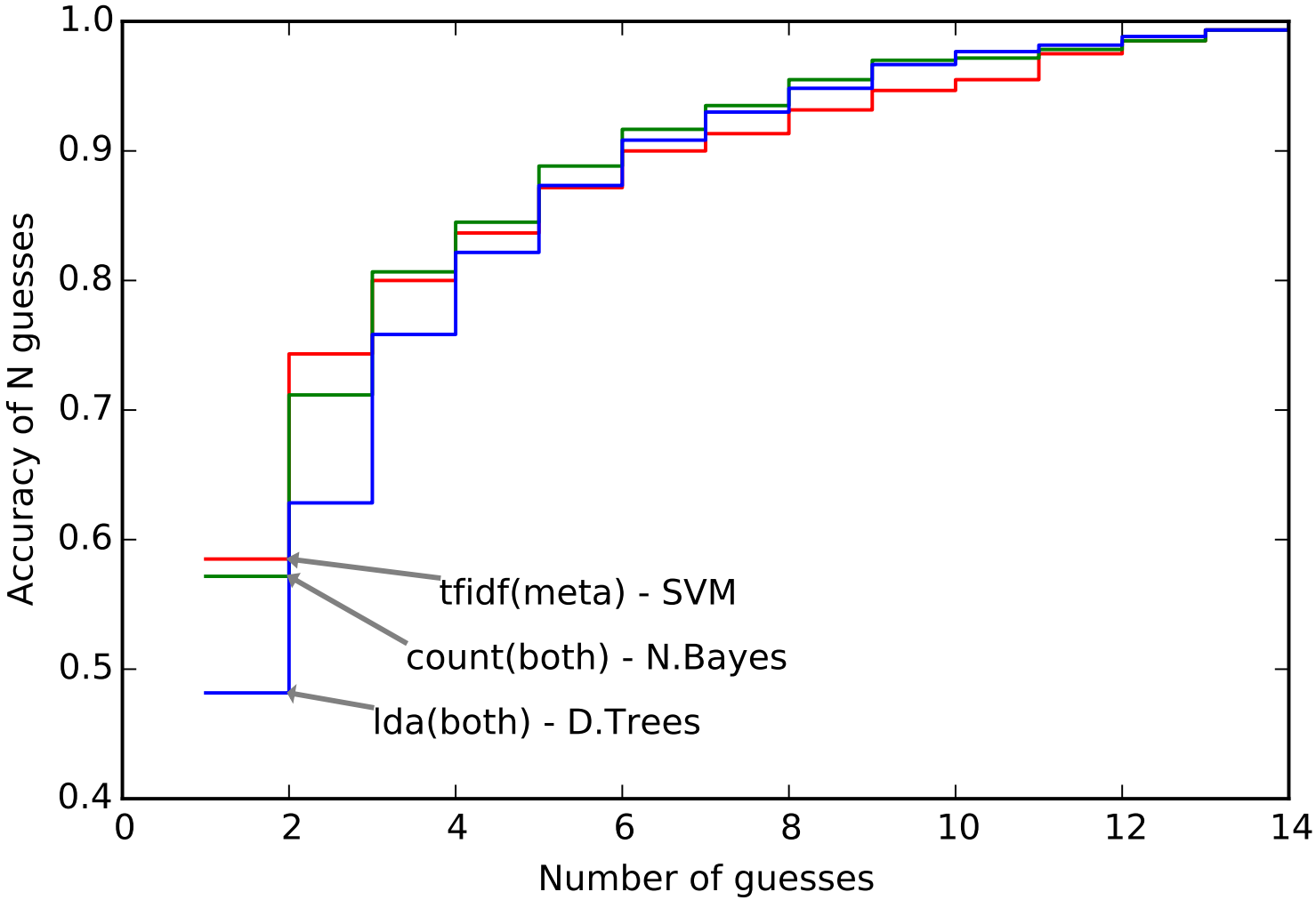
How many guesses does it take to get it right?



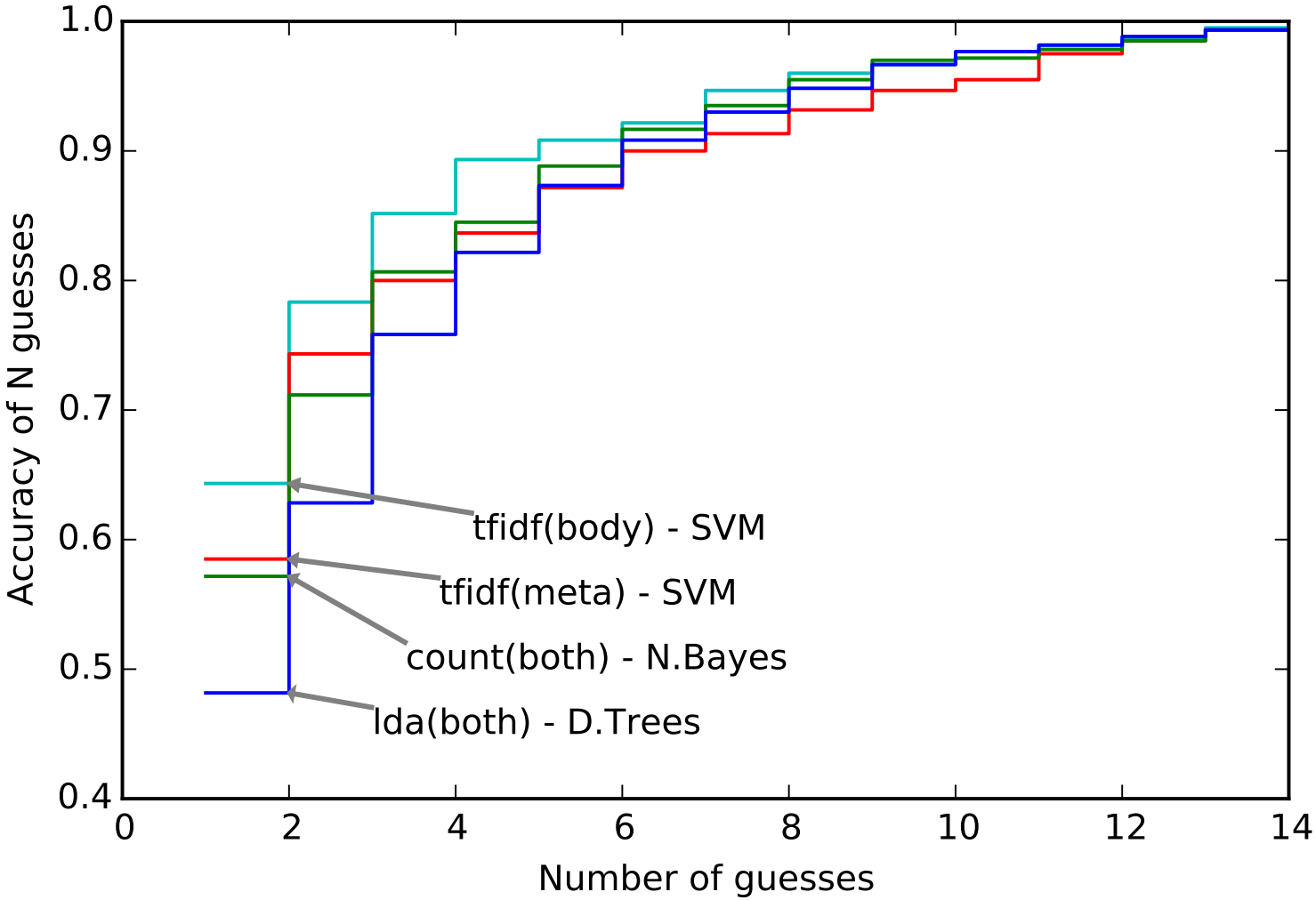
How many guesses does it take to get it right?



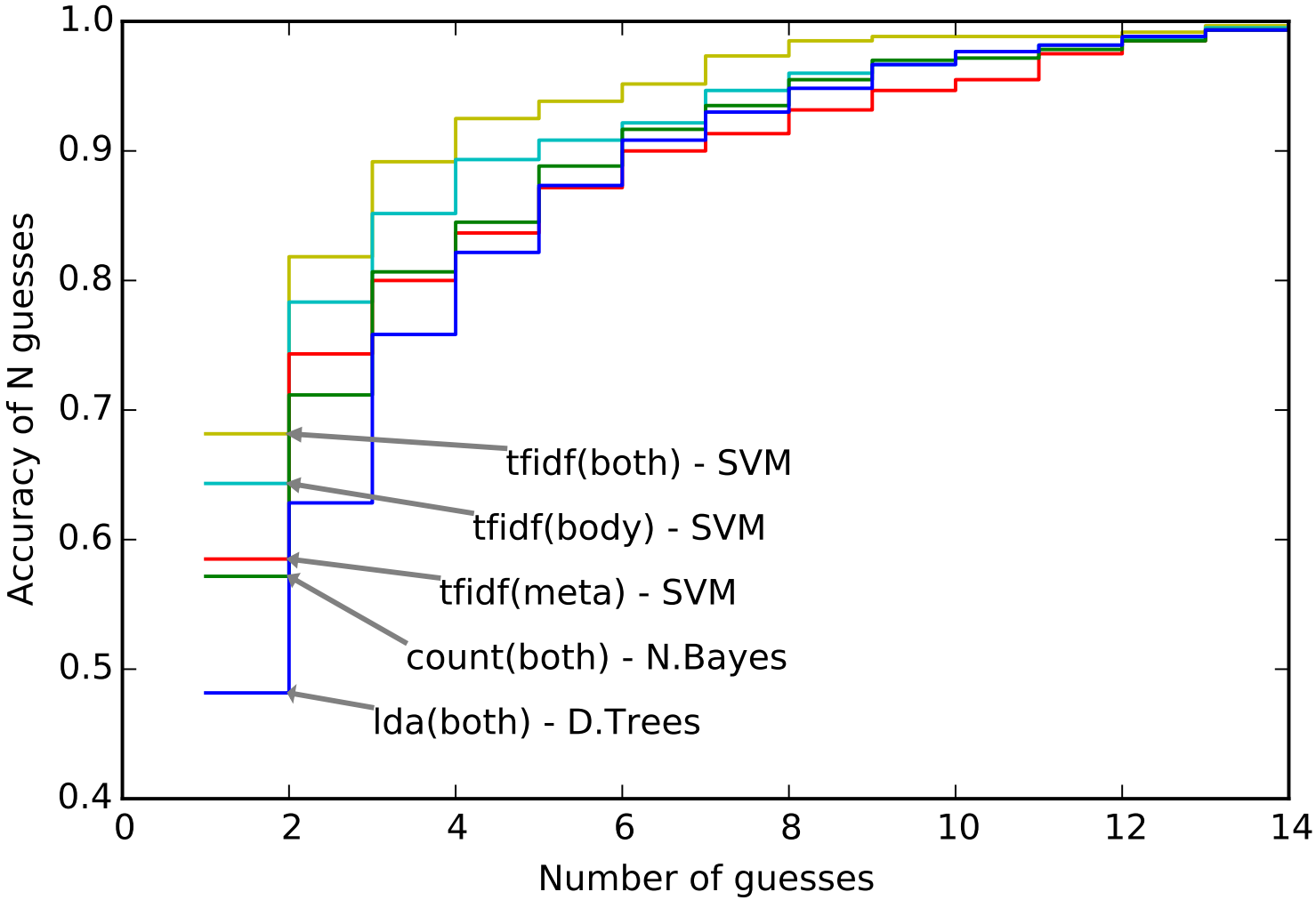
How many guesses does it take to get it right?



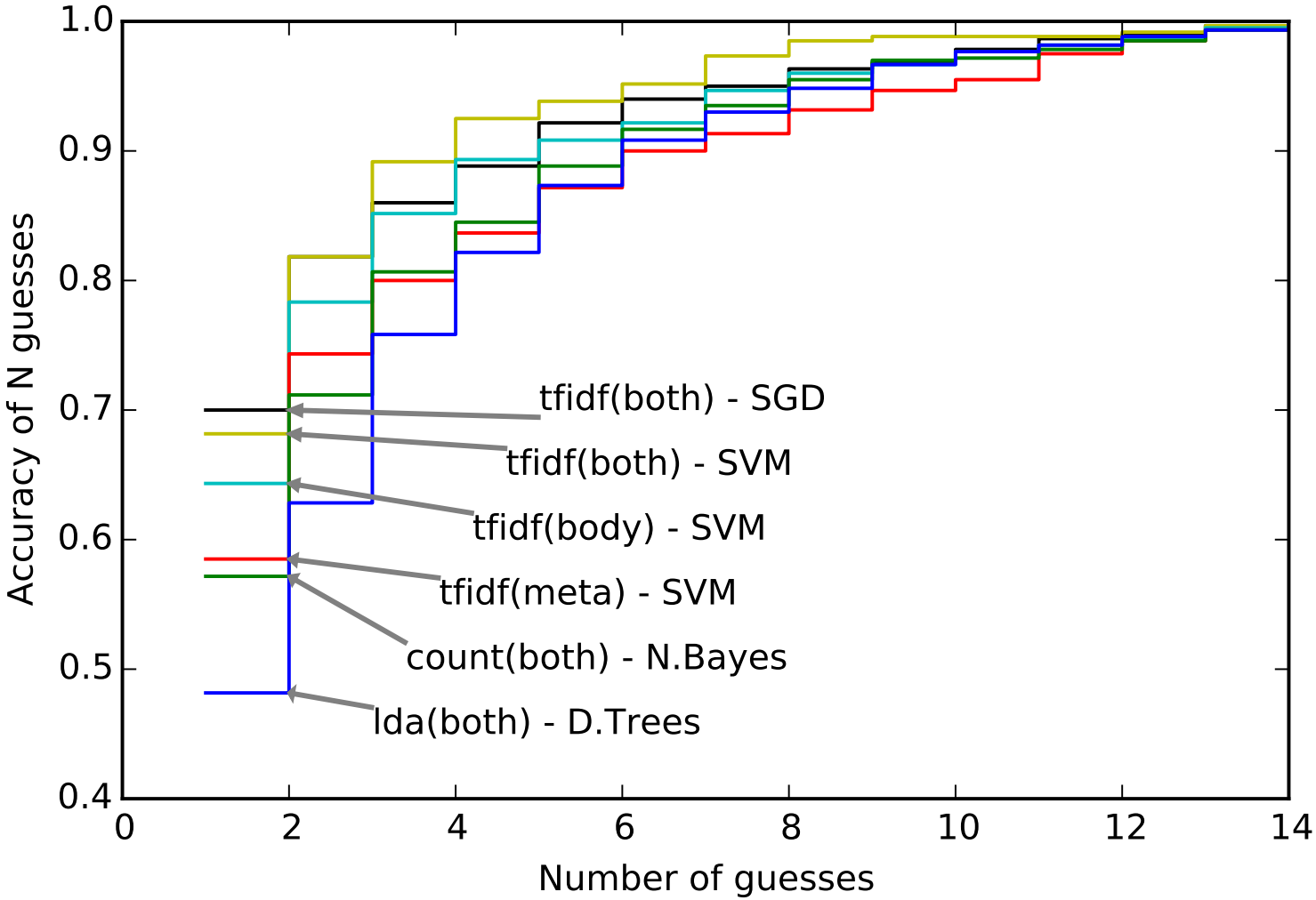
How many guesses does it take to get it right?



How many guesses does it take to get it right?



How many guesses does it take to get it right?



Top Words

SVM feature weightings for website body

1. **arts**: art, artist, award, design, costum, font, film, tattoo
2. **business**: busi, compani, product, investor, drill, sustain, financi, industri
3. **computers**: comput, file, usenet, softwar, bb, algorithm, robot, window
4. **games**: game, dice, puzzl, rpg, wargam, miniatur, stack, diplomaci
5. **health**: health, pharmaci, addict, diseas, drug, osha, age, veterinari
6. **home**: garden, diy, apart, toy, consum, clean, move, kitchen
7. **kids and teens**: kid, parent, teacher, clipart, fun, homeschool, children, disney
8. **news**: newspap, netanyahu, weather, obama, news, radio, iran, polit
9. **recreation**: hike, lock, climb, bird, cave, travel, camp, pet
10. **reference**: librari, knot, dictionari, isbn, thesauru, biographi, word, encyclopedia
11. **regional**: popul, countri, cia, island, nat, territori, geograph, gdp
12. **science**: scienc, astronomi, physic, scientif, econom, laboratori, mathemat, chemistri
13. **shopping**: furnitur, accessori, 00, 99, price, shop, camera, pipe
14. **society**: astrolog, sex, aesthet, horoscop, crime, legend, folklor, gay

Top Words

SVM feature weightings for website body

1. **arts**: art, artist, award, design, costum, font, film, tattoo
2. **business**: busi, compani, product, investor, drill, sustain, financi, industri
3. **computers**: comput, file, usenet, softwar, bb, algorithm, robot, window
4. **games**: game, dice, puzzl, rpg, wargam, miniatur, stack, diplomaci
5. **health**: health, pharmaci, addict, diseas, drug, osha, age, veterinari
6. **home**: garden, diy, apart, toy, consum, clean, move, kitchen
7. **kids and teens**: kid, parent, teacher, clipart, fun, homeschool, children, disney
8. **news**: newspap, netanyahu, weather, obama, news, radio, iran, polit
9. **recreation**: hike, lock, climb, bird, cave, travel, camp, pet
10. **reference**: librari, knot, dictionari, isbn, thesauru, biographi, word, encyclopedia
11. **regional**: popul, countri, cia, island, nat, territori, geograph, gdp
12. **science**: scienc, astronomi, physic, scientif, econom, laboratori, mathemat, chemistri
13. **shopping**: furnitur, accessori, 00, 99, price, shop, camera, pipe
14. **society**: astrolog, sex, aesthet, horoscop, crime, legend, folklor, gay

Pipeline & Tools

- Scikit-learn
  - Python machine learning library
  - Range of classifiers and other processing tools
  - Built-in pipeline framework for cascading algorithms
- Gensim, a python topic modelling toolbox used for LDA
- NTLK, python natural language library used for word stemming

Final Pipeline

1. Sanitise
   - Strip tags
   - Remove stopwords
   - Stem words
2. Tf-idf, both metadata and body
3. Support Vector Classification

Cross-validated accuracy of 56%

NB. a random assignment would give an accuracy of 6%

Limitations

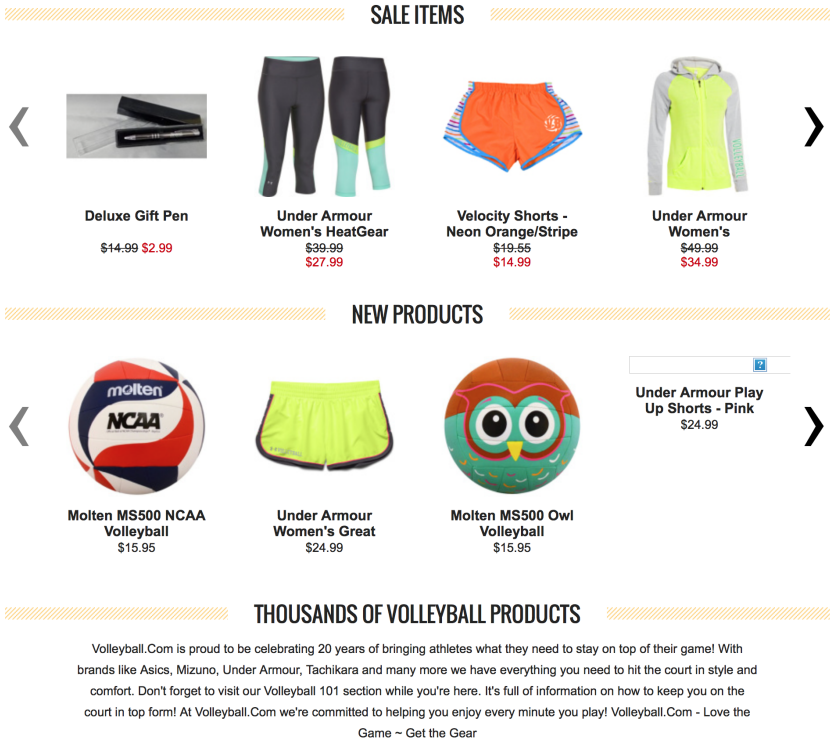
Issues & Difficulties

- Polysemy; multiple meanings of the same word.
- Bag of words model, order is disregarded.
- Different results each iteration due to randomness in Gibbs Sampling.
- Stemming, capturing the meaning of words. Literary tenses.
- Topics in a document can be subjective not objective.
- Quantifying performance, critical evaluation.
- Storing and processing large quantities of corpora.
- Currently using old websites, which tend to lack content and metadata.

Multiple correct answers

Misclassifications

DMOZ class was Sports, top two guesses were Shopping, Sports

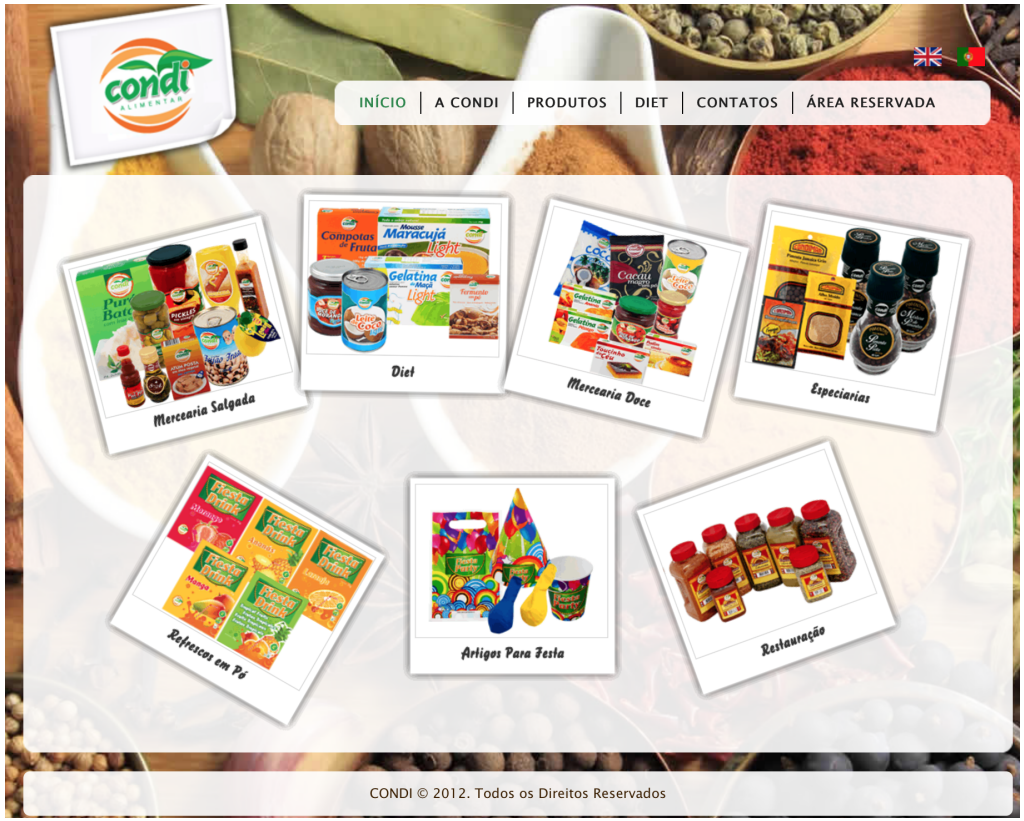


This is clearly a sports shopping website!

Training Bias

Misclassifications

DMOZ class was Business, misclassified as Sports

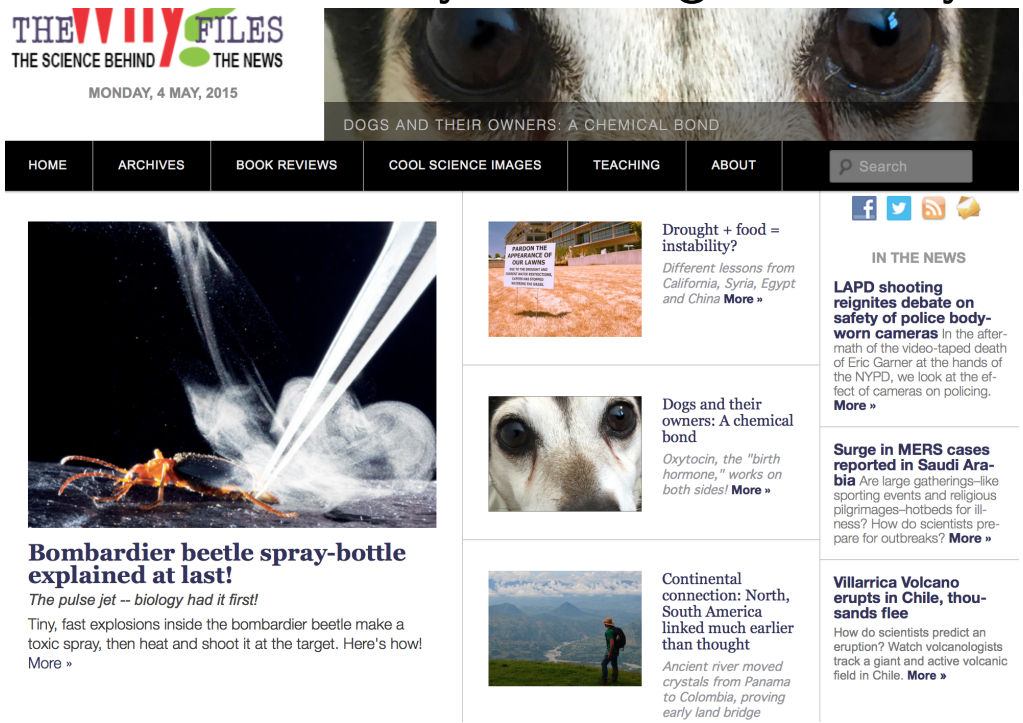


The website has practically no text or metadata.

Objective nature of the problem

Misclassifications

DMOZ class was "Society", our algorithm says Science.



We think DMOZ is wrong here! Many 'society' websites on DMOZ seem to be about other things.

Extensions

Future work . . .

- Learn and classify a topic heirarchy
- Deal with (almost) empty websites
  - Filter them out entirely?
  - Train a 'null' category?
- Recognise websites not in any existing category
- Try other classifiers (e.g. Neural Networks)
- Make a website service

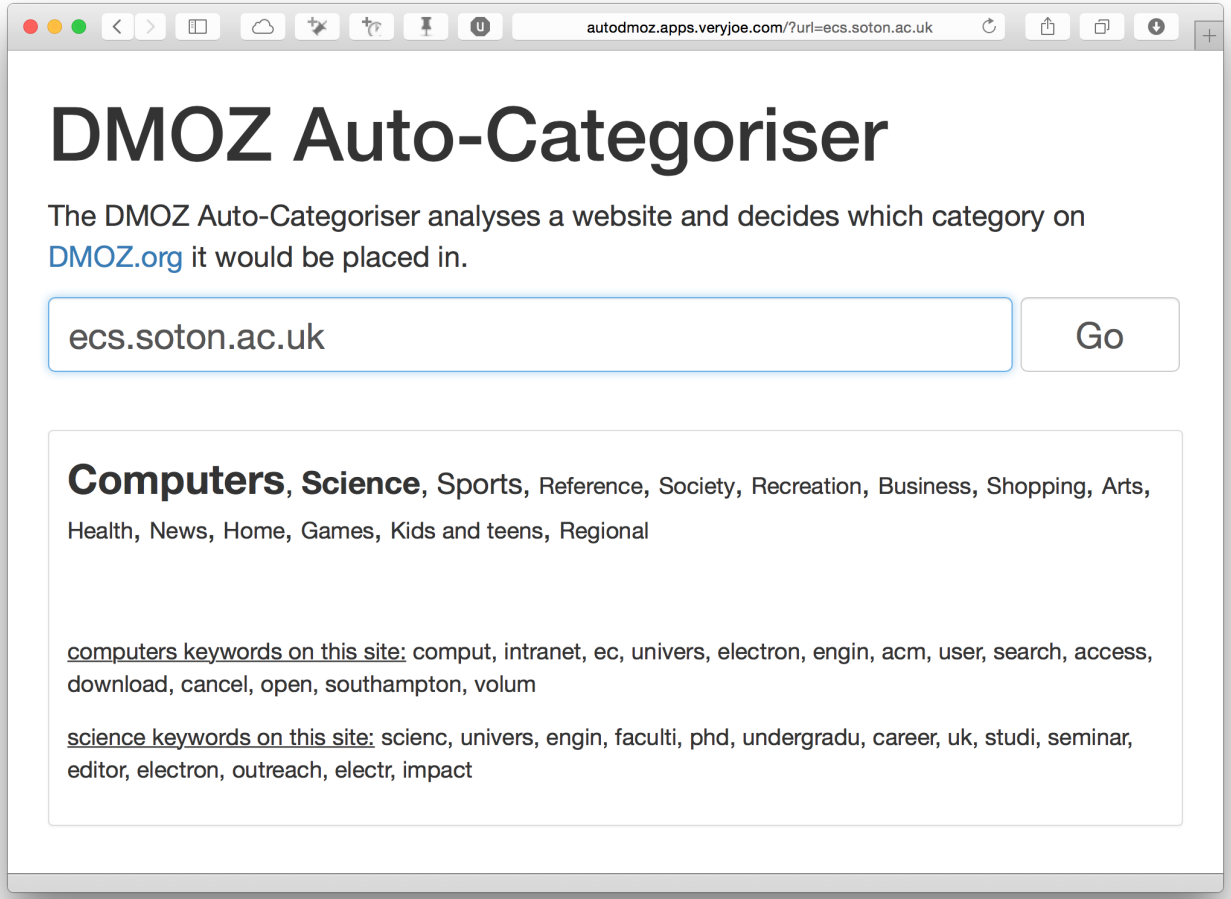
Extensions

Future work . . .

- Learn and classify a topic heirarchy
- Deal with (almost) empty websites
  - Filter them out entirely?
  - Train a 'null' category?
- Recognise websites not in any existing category
- Try other classifiers (e.g. Neural Networks)
- Make a website service
  - Oh wait, we did that!

Web service!

`http://autodmoz.apps.veryjoe.com`



The End

Any questions?