# Understanding Informal Communication

Jason Rennie
jrennie@csail.mit.edu

Joint work with Tommi Jaakkola

# What Is Informal Communication?

- E-mail
- Bulletin Boards
- Mailing Lists
- Newsgroups

# Extracting Information

- Q: What restaurants are in the South End?
- A: Deluxx Cafe (among others)

SUBJECT: A Trip to Fenway etc.

...

4) **Deluxx Cafe** in the **South End** is kitzch or dive depending on your pov. Great food fairly cheap with an attractive mix of clientelle.

...

## Up-to-Date Information

- Traditional sources of Boston restaurant information:
  - Zagat's
  - Boston.com
- Informal Communication provides more up-to-date information.

## Restaurant Bulletin Board Post

SUBJECT: Restuarant soon to be formerly known as Bickfords.
Hi All,
I just got back from what will be the soon to be newly renamed and already revamped Bickfords in Framingham- for the second time. The name change is coming but I don't know when, it will have the word "tavern" in it is all I was told.
The interior is basically the same but with the addition of dark wood and divider panels.
The new menu has traditional Bickfords fare with the addition of some new items. The new items are seafood and from talking to the staff, seafood will be a major menu item.

## Restaurant Tasks

- What new restaurants opened in Boston in the last month?
- If I like Legal's, Emma's and Central Kitchen, what other restaurants would I like?
- What dishes do people like at Atasca?

## Restaurant Sub-Tasks

- Identifying restaurant names
- Resolving references to restaurants
- Tracking context (what restaurant is this person talking about?)
- Determining sentiment (is the review positive or negative?)
- Collaborative filtering

## Informal Communication Hard to Understand

SUBJECT: Boozehound. Outside drinks tonight?

Unfortunately, bukowski's isn't nearly as dark as it should be. I went in for a drink Saturday while waiting for a table at ECG. Great beer's but I was hoping for the ambiance of the boston location. The Inman square one is nice, just a little different.

## Named Entity Detection

- Named Entities are people, places, locations, company names, restaurants, etc.
- Easy to detect in formal communication
- 95% accuracy common

## Detecting Restaurant Names

- Most unusual restaurant name: “Cambridge, 1”

SUBJECT: Quiet But Good?

I want to go out to a new top-end restaurant. I’ve been to many of the good ones, i.e., **Radius**, **L’Espalier**, **Locke-Ober**, **Biba**, **Julien**, **Aujourd’hui**, **Hamersley’s**, **Rialto**, **Mama Maria**, **Maison Robert**, **Harvest** and of course all the steakhouses.
I’m thinking of: **Clio**, **#9 Park**, **Upstairs on the Square**, **Oleana**, **Troquet**, **Federalist**. None of which I’ve been to.

## New Features Needed

SUBJECT: Death, Custard & Fond Memories, Wisconsin Style.

Yup, i've had that wisconsin custard . and in fact it's great & unlike any soft serve. My understanding (from the custard folks) was that it has a #$%load of egg yolks in it that makes it so rich . I once smuggled a loved one . a butterscotch sundae version of same . into an intensive care unit at a Milwaukee hosital. He died a few weels later . but the smile . on his face . after weeks of hospital gruel . will always be . a fond (du lacian) & bonding memory for us both. by the way don't bemoan the fact the stuff isn't imported into new england, Its one of the few good reasons to go back.

## New Features Needed

SUBJECT: **finale** harvard square.

has anyone been to the recently opened **finale** in harvard square? i've walked by a few times and noticed that they offer (a pretty busy) lunch service as well - sandwiches and items from the bakery case. any input? thanks in advance.

## Named Entities are Topic-Centric

- Named Entities are a common topic in news articles, discussions, etc.
- Restaurant names often *the* topic of discussion.

## Inverse Document Frequency

- Common words get low score
- Rare words get high score

$$\mathrm{IDF}(w) = \log \frac{(\text{total \# of documents})}{(\text{\# documents with } w)} \tag{1}$$

## The Unigram

- Simplest model of word frequency
- Assumes word occurrences are independent

$$P(\vec{w}) = \prod_i \theta_{w_i} \tag{2}$$

## Topic-Centric Distribution

- Two modes:
  - Low/Off-Topic: word might be mentioned in passing; near zero chance of occurrence
  - High/On-Topic: word mentioned repeatedly

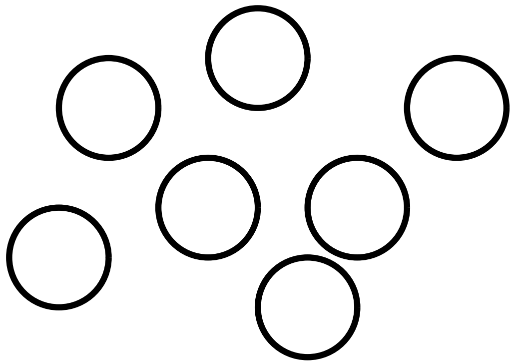
## Identifying Topic-Centric Words

- "Sichuan" occurs in 3 out of 132 threads
- Overall rate: .001 (1/1000)

| Thread | 1 | 2 | 3 |
|---|---|---|---|
| Occurrences | 23/856 | 28/3193 | 1/1965 |
| Rate | .027 | .0088 | .0005 |

# How to Identify: Intuition

- Non-topic word: empirical rates cluster



- Topic word: empirical rates spread out

## How to Identify: Technical Details

- We use the log-odds ratio of mixture and unigram models:

$$s_{\text{mix}} = \log \frac{p_{\text{mix}}(\vec{h}, \vec{n}; \lambda, \phi_1, \phi_2)}{p_{\text{uni}}(\vec{h}, \vec{n}; \theta)} \tag{3}$$

- We call this the “Mixture Score”.

## Restaurant Data

SUBJECT: Any **Speed** sightings?

I saw the report of a rumor a few days ago on the board that **Speed**'s hot dogs had closed up shop due to a fire, so I went to check it out today, and there was no sign of his truck. [...]

SUBJECT: My week in food: Skipjack's, Chacarero, Turner Fisheries, Excelsior.

[...] I think it has taken some time since they reopened to get up to **speed**, but I think Turner Fisheries is now a good alternative. [...]

## Words Ranked by Mixture Score

| Token | Score | Restaurant |
|---|---|---|
| **sichuan** | 99.62 | 31/52 |
| **fish** | 50.59 | 7/73 |
| was | 48.79 | 0/483 |
| **speed** | 44.69 | 16/19 |
| **tacos** | 43.77 | 4/19 |
| **indian** | 41.38 | 3/30 |
| sour | 40.93 | 0/31 |
| **villa** | 40.36 | 10/11 |
| **tokyo** | 39.27 | 7/11 |
| greek | 38.15 | 0/20 |

## Top Restaurant Names (50%+ usage)

| Rank | Token | Restaurant |
|---|---|---|
| 1 | sichuan | 31/52 |
| 4 | speed | 16/19 |
| 8 | villa | 10/11 |
| 9 | tokyo | 7/11 |
| 21 | zoe | 10/11 |
| 22 | penang | 7/9 |
| 23 | pearl | 11/13 |
| 26 | dhaba | 8/13 |
| 29 | gourmet | 23/27 |
| 30 | atasca | 9/10 |

## Mixture Score of An Average Restaurant Token

| Score | Average Score | Median Score |
|---|---|---|
| Mixture | **2.5** | 2.9 |
| Baseline | 1.0 | 1.0 |

## Mixture Score Faults

- Author style
- Type of post

## Pros and Cons

| | Mixture Score | IDF |
|---|---|---|
| Pro | Some topic words get very high score | Common words get very low score |
| Con | High score to author, style tendencies | Can't discriminate medium to low frequency words |

## Scores of An Average Restaurant Token

| Score | Average Score | Median Score |
|---|---|---|
| Mixture | **2.5** | 2.9 |
| IDF | 1.9 | **3.90** |
| Baseline | 1.0 | 1.0 |

## Are They Independent?

| Condition | Restaurant |
|---|---|
| Mixture $> 4.0$ | 176/325 |
| IDF $> 4.0$ | 170/325 |
| Mix $> 4.0$ and IDF $> 4.0$ | 93/325 |

## IDF*Mixture Score

- Multiplication is relaxation of AND operator.

# Top IDF*Mixture Words

| Token | Score | Restaurant |
|---|---|---|
| **sichuan** | 376.97 | 31/52 |
| **villa** | 197.08 | 10/11 |
| **tokyo** | 191.72 | 7/11 |
| ribs | 181.57 | 0/13 |
| **speed** | 156.25 | 16/19 |
| **penang** | 156.23 | 7/9 |
| **tacos** | 153.05 | 4/19 |
| **taco** | 138.38 | 1/15 |
| **zoe** | 134.23 | 10/11 |
| festival | 127.39 | 0/14 |

| Rank | Token | Restaurant |
|---|---|---|
| 1 | sichuan | 31/52 |
| 2 | villa | 10/11 |
| 3 | tokyo | 7/11 |
| 5 | speed | 16/19 |
| 6 | penang | 7/9 |
| 9 | zoe | 10/11 |
| 12 | denise | 5/8 |
| 16 | pearl | 11/13 |
| 19 | khao | 4/7 |
| 21 | atasca | 9/10 |
| 23 | bombay | 6/7 |

# The IDF*Mixture Score

| Score | Avg. Score | Med. Score |
|---|---|---|
| IDF*Mixture | **7.20** | **17.15** |
| $\text{Mixture}^2$ | 4.61 | 8.35 |
| $\text{IDF}^2$ | 2.31 | 15.19 |

## Named Entity Detection

- Use a standard set of NED features (capitalization, part-of-speech, punctuation, etc.)
- Include Mixture Score and IDF features.
- Evaluate using F1-breakeven (higher is better).

## NED: Results

| Features | F1 brkevn |
|---|---|
| Baseline | 55% |
| IDF | 56% |
| Mixture | 56% |
| IDF,Mix | 57% |
| IDF*Mixture | **59%** |

## Result Significance

- Wilcoxon signed rank test: IDF*Mixture result significant at 5% level

## Summary

- We introduced a new informativeness criterion (Mixture score).
- We showed that words with high Mixture scores tend to be informative.
- We showed that, especially in conjunction with IDF, the Mixture score is highly effective at identifying named entities (restaurants).