Text Mining Analysis: Data-focused job listings

Kimberley Mitchell
http:/mitchki.com

Project Goals

Detect distinctions between job postings

1. between data-related keywords
2. between major cities

(classification modeling)

Identify latent topics (topic modeling)

Factors investigated

Job type keywords:

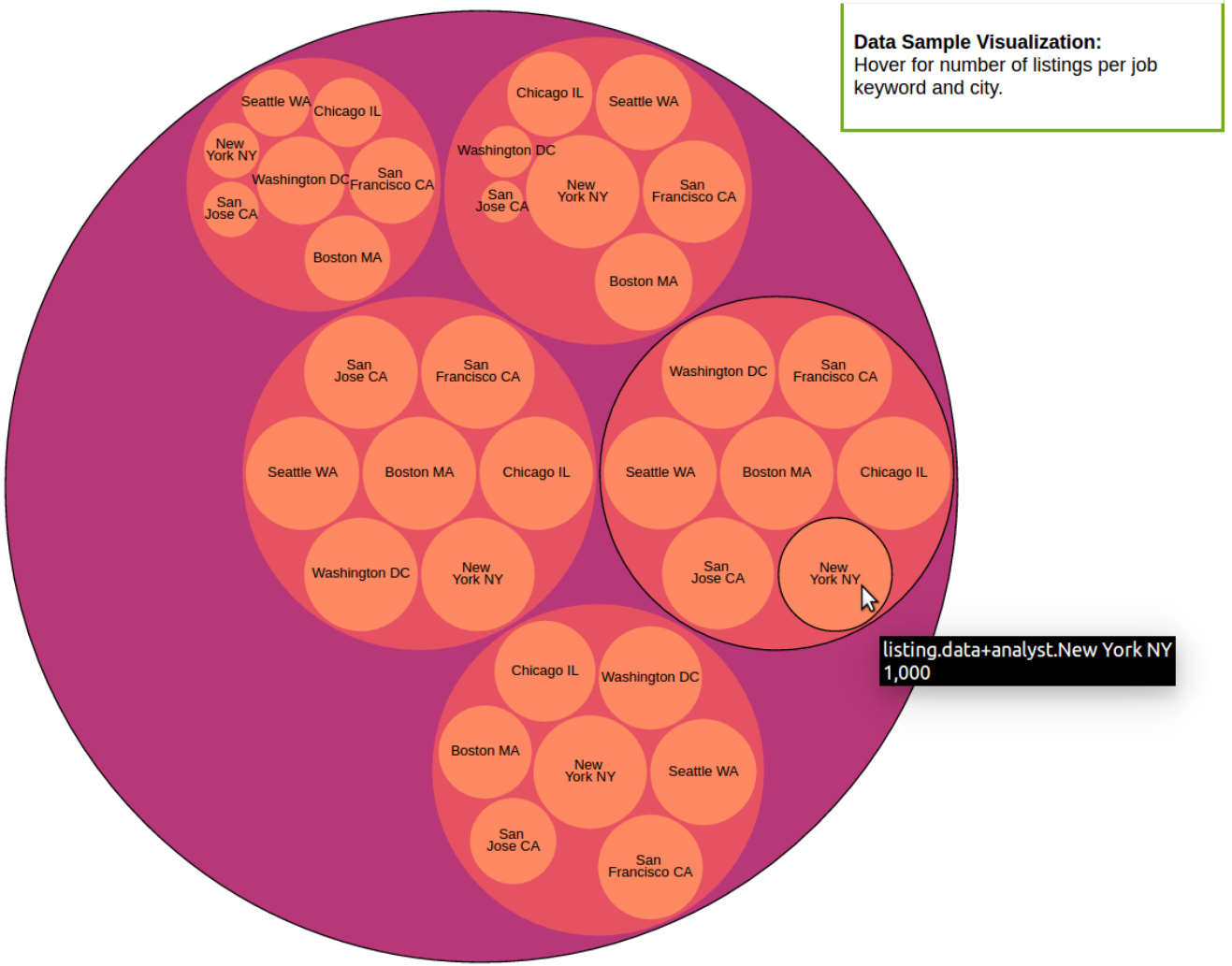
- Data scientist
- Data engineer
- Data architect
- Data analyst
- Statistician

Cities

- New York, NY
- San Francisco, CA
- San Jose, CA
- Boston, MA
- Seattle, WA
- Chicago, IL
- Washington, DC

Data Sample Visualization

http://mitchki.com/D3/joblist.html



Methodology

- Data
  - Use indeed.com API to request 1000 URLs for each combination of job type and city
  - Scrape each URL to obtain job listings, store in MongoDB
  - Clean, tokenize, and vectorize; split into test and train

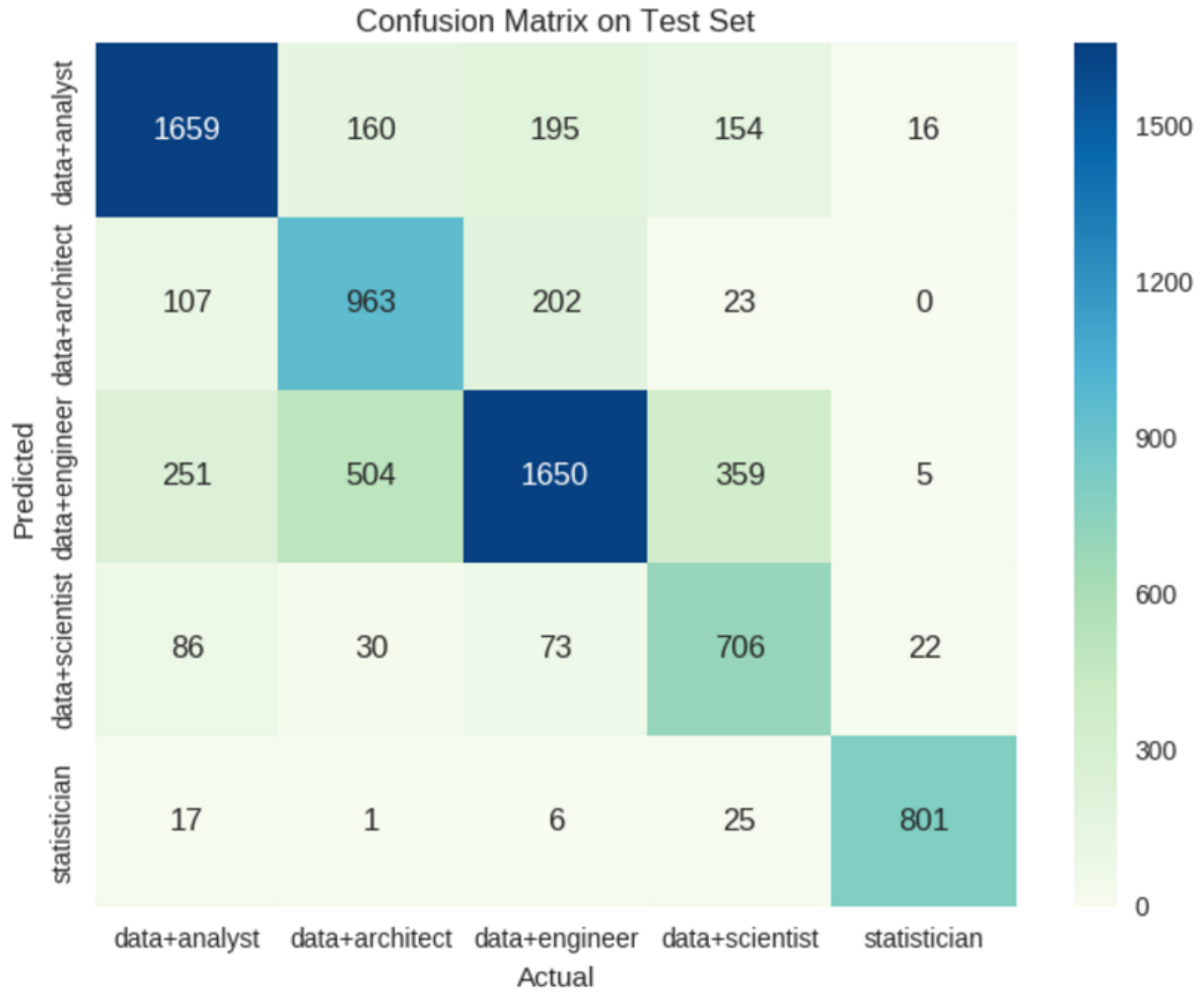
- Modeling
  - Multinomial Naive Bayes to distinguish among job types, then drill down to distinguish cities
    - Precision/recall confusion matrix to measure models
  - Latent Dirichlet Allocation (LDA) to extract latent topics

Results: Job types (test set)

	precision	recall	f1-score	support
data+analyst	0.76	0.78	0.77	2120
data+architect	0.74	0.58	0.65	1658
data+engineer	0.60	0.78	0.67	2126
data+scientist	0.77	0.56	0.65	1267
statistician	0.94	0.95	0.95	844
avg / total	0.73	0.72	0.72	8015

Test Set Accuracy: 0.76

Confusion Matrix: Job types



Unique features among top 50 per job type

- **Data analyst:** 'analyst', 'communication', 'financial', 'marketing', 'process', 'reporting', 'responsibilities'
- **Data architect:** 'application', 'architecture', 'client', 'cloud', 'enterprise', 'understanding', 'web'
- **Data engineer:** 'build', 'building', 'computer', 'engineer', 'engineers', 'test'
- **Data scientist:** 'care', 'center', 'health', 'learning', 'medical', 'scientists'
- **Statistician:** climate', 'excellent', 'field', 'lead', 'manage', 'programming', 'quality', 'quantitative', 'sas', 'statistical', 'study'

Common terms among top 50 per keyword

Intersection of all 5 keywords:

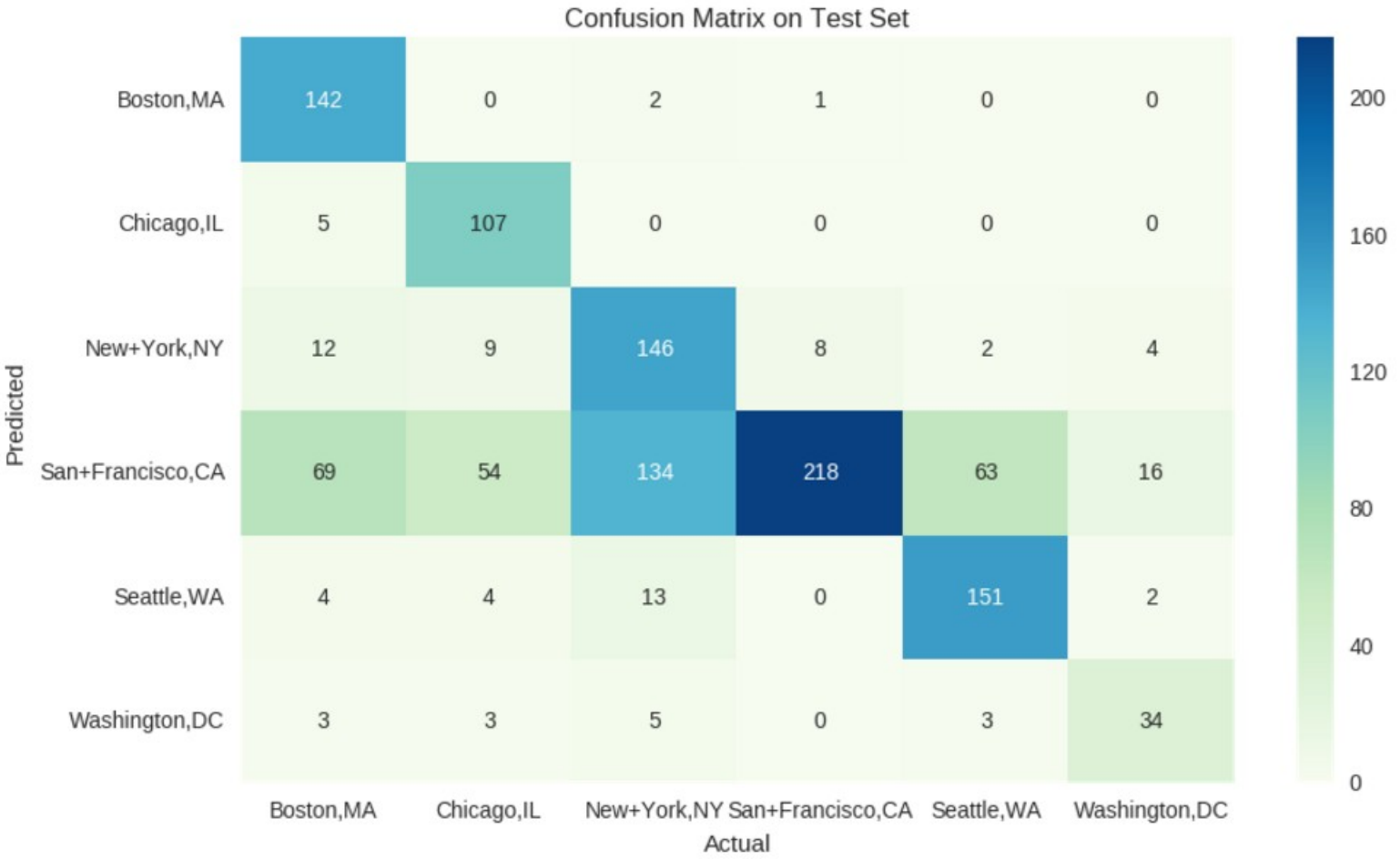
- ability', 'business', 'development', 'environment', 'including', 'information', 'knowledge', 'management', 'product', 'project', 'required', 'services', 'skills', 'strong', 'team', 'technical', 'the', 'we', 'work', 'working', 'years', 'you'

Results: cities

	precision	recall	f1-score	support
Boston,MA	0.98	0.60	0.75	235
Chicago,IL	0.96	0.60	0.74	177
New+York,NY	0.81	0.49	0.61	300
San+Francisco,CA	0.39	0.96	0.56	227
Seattle,WA	0.87	0.69	0.77	219
Washington,DC	0.71	0.61	0.65	56
avg / total	0.79	0.66	0.68	1214

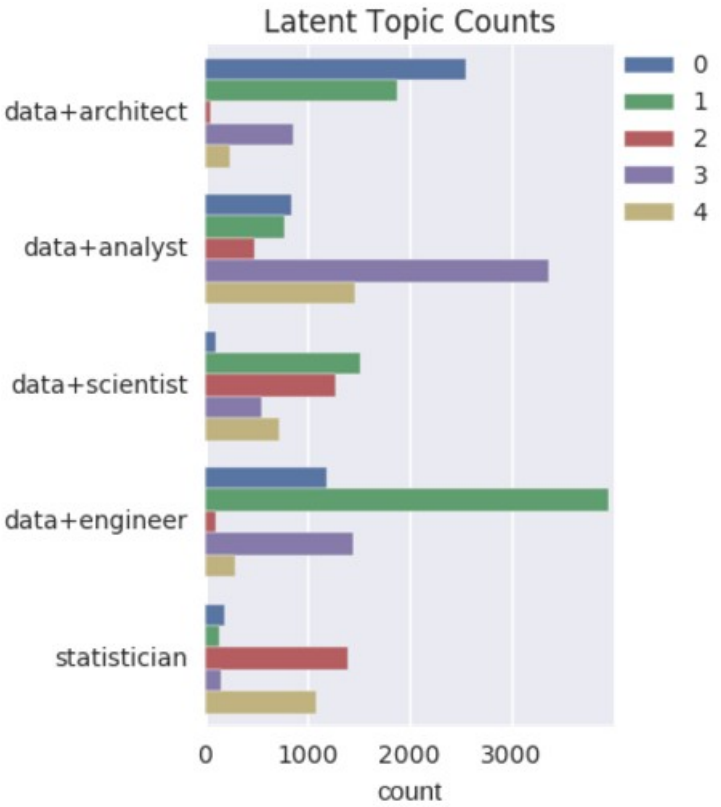
Accuracy Score(Test): 0.65

Confusion Matrix: cities



LDA latent topics (n=5)

- Topic 0 - technical, solutions, design, technology, management, security, requirements, knowledge, systems, support
- Topic 1 - you, software, your, new, design, systems, engineering, science, company, from
- Topic 2 - research, clinical, health, medical, statistical, care, analysis, staff, all, including
- Topic 3 - management, support, all, information, project, required, by, including, job, systems
- Topic 4 - sales, you, product, marketing, analytics, customer, analysis, products, strong, teams



Questions?

Kimberley Mitchell
kim@who-knows.com