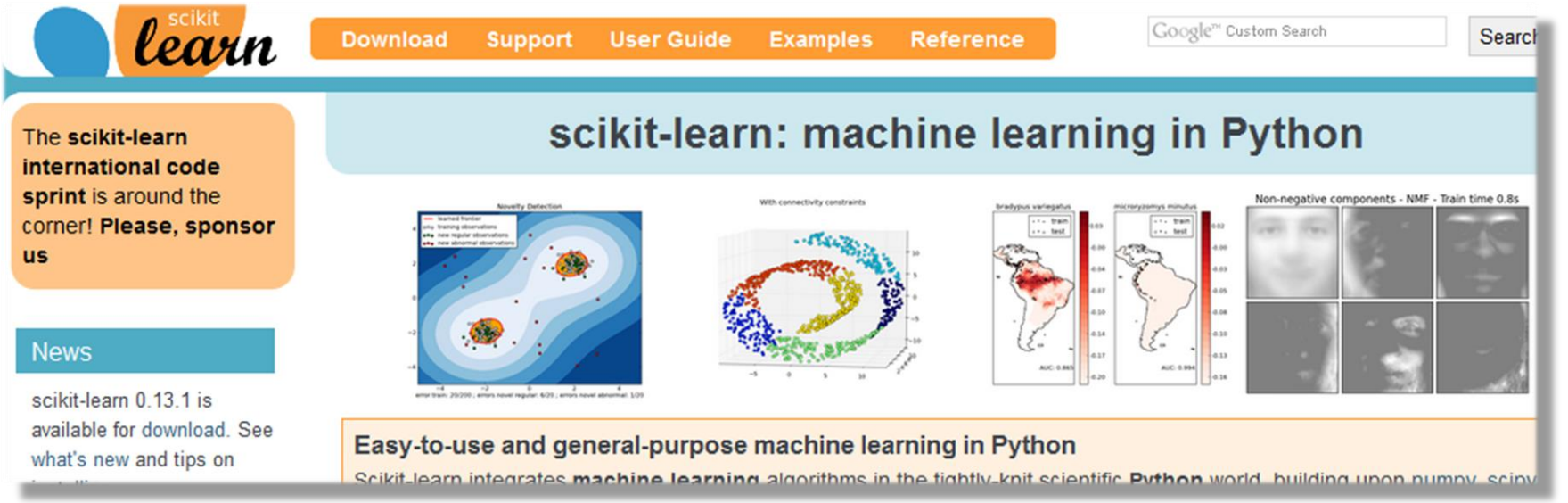
Deliverable #4

Marie-Renée Arend
Josh Cason
Anthony Gentile

4 June 2013

Big idea: Classification

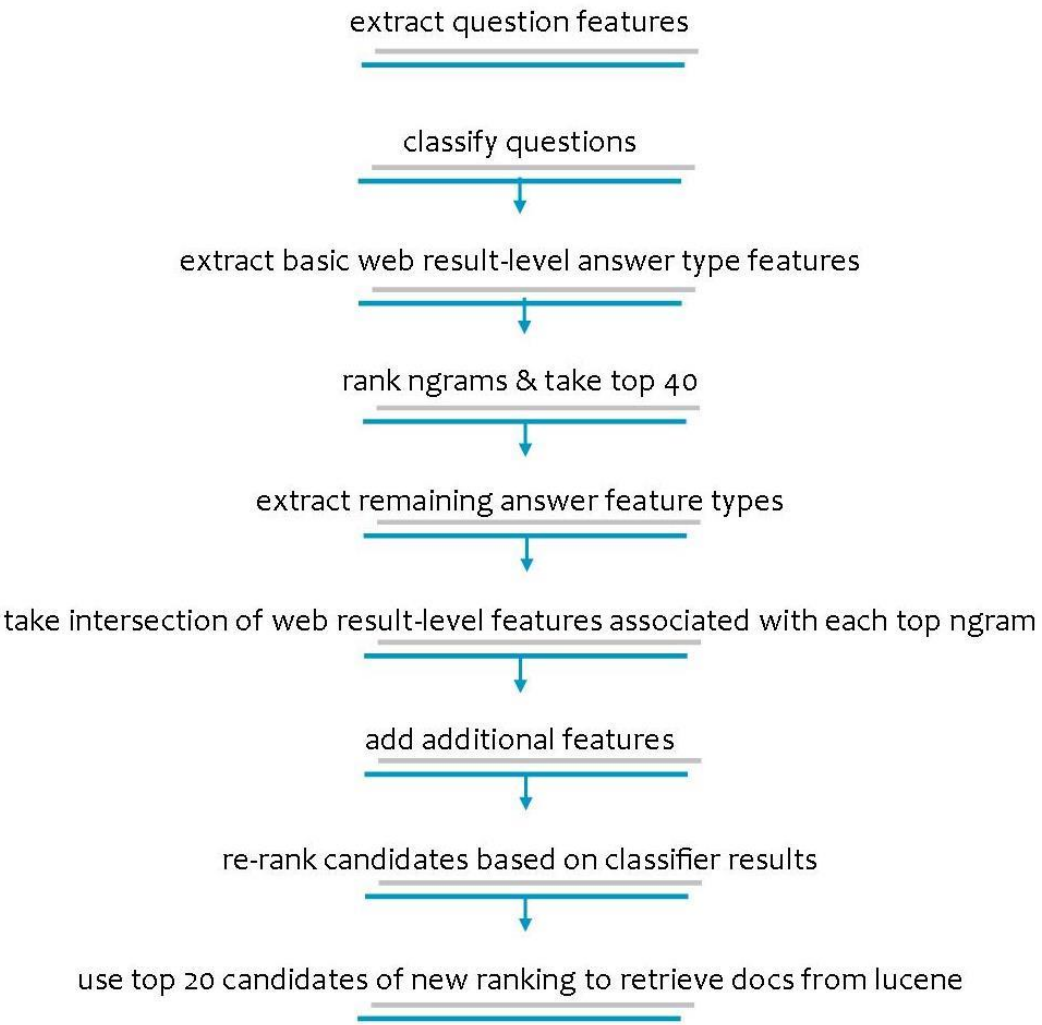


- Scikit Learn python package
- Support Vector Machines classifier (Radial basis function kernel)
- Chi Squared feature selection

Big Idea: Caching

- Everything.

System Pipeline



Query Processing

- Approaches tried in previous versions:
  - D2: basic shallow processing
  - D3: using lexical resources
- Classifier approach:
  - D4: *loosely based on Li & Roth's syntactic features*
    - Stemmed ngrams (*n* = 1,2,3,4)
    - Weights for temporal, location or numerical question words
    - POS-tagged tokens from question & target with stopwords removed
    - Head NP & VP chunks – handwritten grammar
    - Question word(s)
  - *Issues:*
    - Addition of extra features beyond unigrams didn't make a significant difference & increased total runtime
    - Final system: features are unigrams

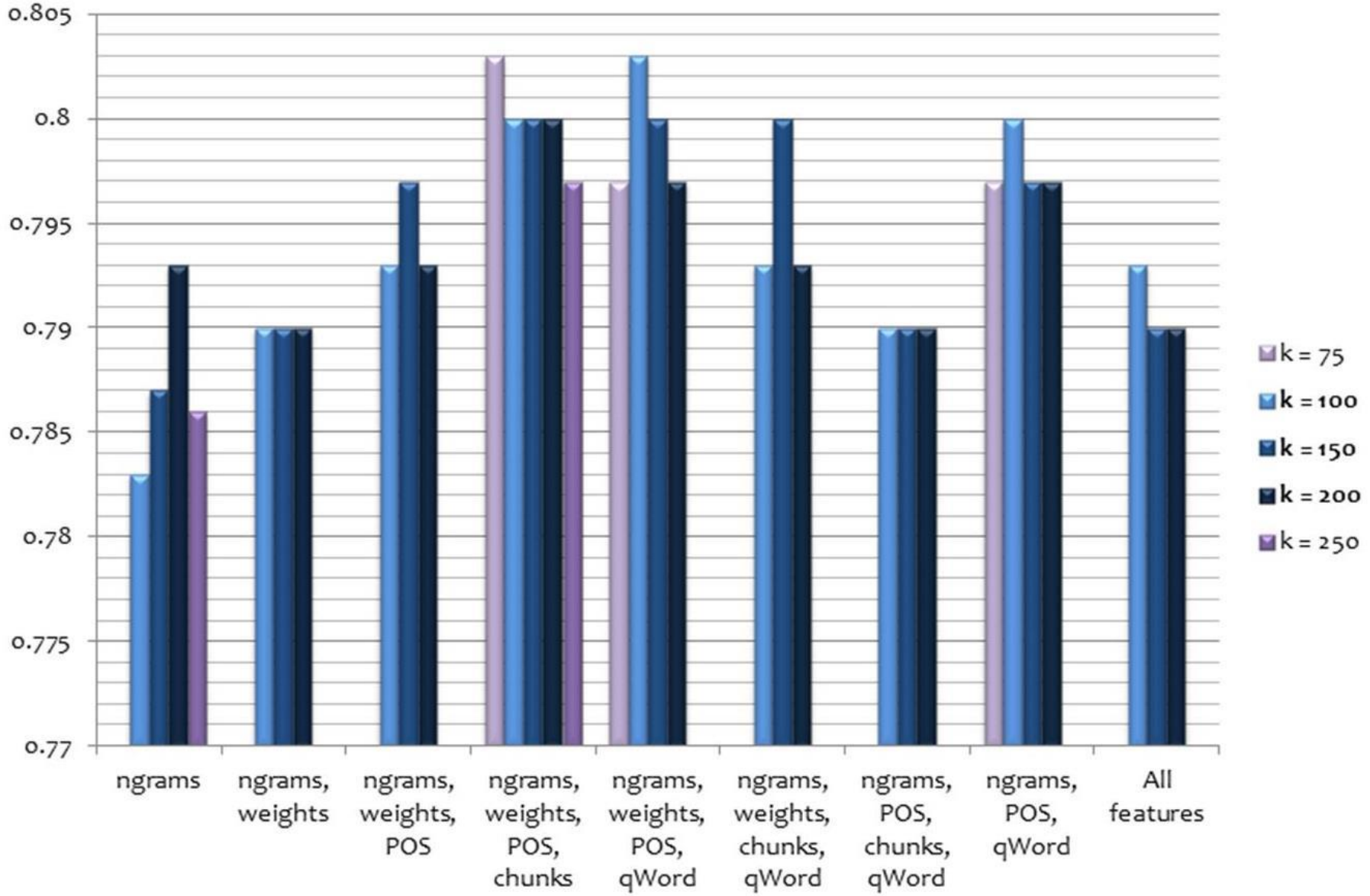


Fig. 1: Features and Performance *(experimentation phase)*

Classifier & Web-based Boosting

- Train question classifier (qc)
- Classify question
- Extract web result-level answer type features that require punctuation guided by qc
  - Before text processing a web result
  - take the qc, e.g., ABBR
  - extract all punctuation dependent ABBR patterns
  - ABBR_PUNC_ABREV = '(M\.D\.|M\.A\.|M\.S\.|A\.D\.|B\.C\.|B\.S\.|Ph\.D|D\.C\.|NAACP|AARP|NASA|NATO|UNICEF|U\.S\.|USMC|USAF|USSR|YMCA)'

Classifier & Web-based Boosting

- Tokenize, remove punct., etc
- Re-rank ngrams & take top 40
  - Use Lin's web redundancy algorithm for re-ranking
- Extract ngram level answer pattern features as guided by qc
  - Similar to above but based on a particular answer candidate – no punctuation patterns
    - *(more info below)*

Classifier & Web-based Boosting

- Add the intersection of all web result-level features associated with each top-40 ngram, n
  - $\bigcap_{w \in W} f(n, w)$
  - Where f returns the set of features for w if n appeared there
- Add additional features like top web result rank

Classifier & Web-based Boosting

- Re-rank based on classifier
  - Each candidate is assigned a probability of being a “yes” answer
  - Training based on checking 2004, 2005 answer candidates against their answer patterns using same features
- Use the top 20 candidates from the new ranking to retrieve docs using lucene

Answer Pattern Detection

We used a set of regular expressions to detect answer types in addition to our existing filters and weighting logic.

If we have a question classified as type:

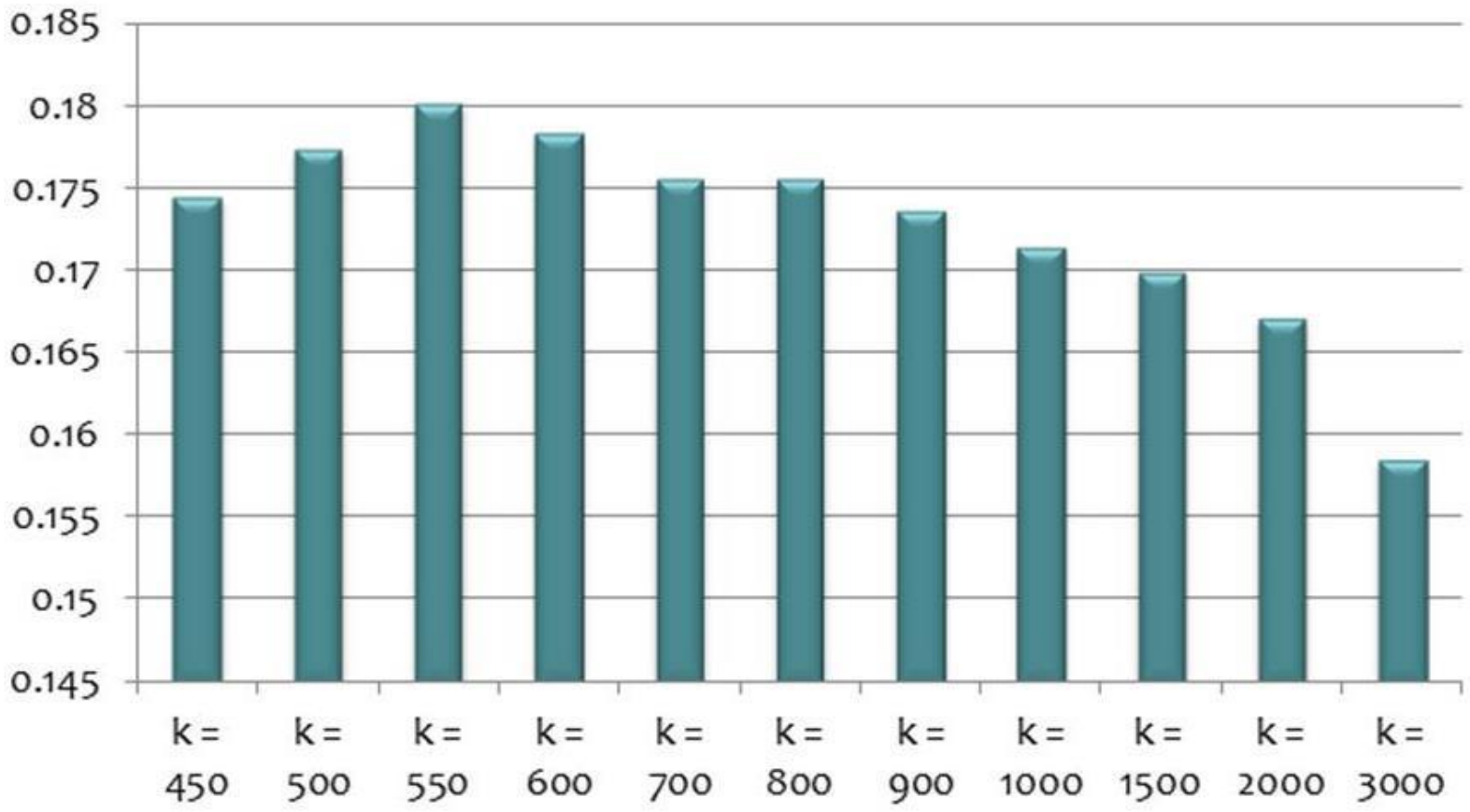
```
['LOC', 'HUM', 'NUM', 'ABBR', 'ENTY', 'DESC']
```

If 'ENTY' , a set of regular expressions for subclasses are triggered (sports, religion, colors, etc):

Example:

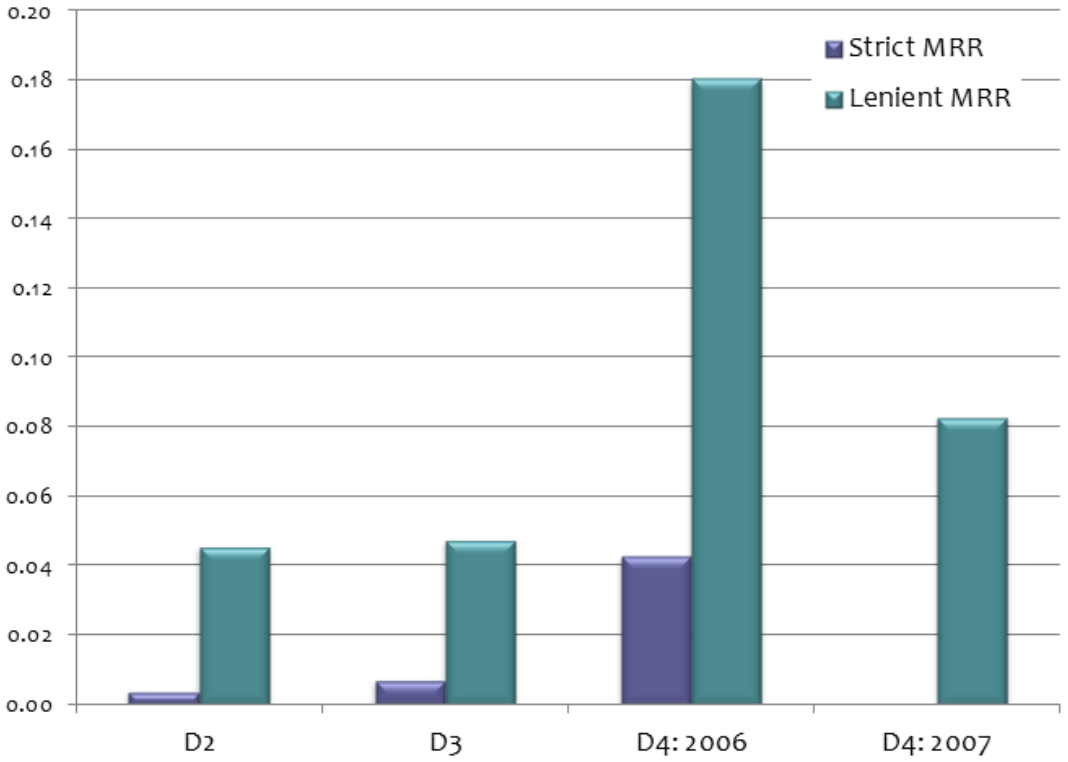
```
ENTY_PLANTS =
set(['rose','weed','tulip','daisy','flower','orchid','bonzai','dog
wood'])
pattern_values['plant'] = ['(' + '|'.join(self.ENTY_PLANTS) + ')']
```

This pattern dictionary is iterated over to find matches in the text and provide for features and boost in weighting for the web results.



Experiment: Select k best features using X^2 selection
(Numbers are lenient MRR scores for 2006)

Results, Issues & Successes



	Strict MRR	Lenient MRR
D2	0.0039	0.0455
D3	0.0070	0.0475
D4: 2006	***0.0426***	***0.1801***
D4: 2007	0.0000	0.0828

- Results analysis
- Issues
  - 0 for 2007 strict MRR
- Successes

- Notes:
  - All answer candidates were less than or equal to 100 chars

Resources

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

Graff, D. (Ed.). (2002). *The AQUAINT corpus of English news text*. Linguistic Data Consortium.

Hatcher, E., Gospodnetic, O., & McCandless, M. (2004). Lucene in action.

Li, X. & Roth, D. (2005). Learning question classifiers: The role of semantic information. *Natural Language Engineering, 1*(1), Retrieved from http://12.cs.uiuc.edu

Lin, J. (2007). An exploration of the principles underlying redundancy-based factoid question answering. *ACM Transactions on Information Systems (TOIS)*,*25*(2), 6.

Mishne, G. & de Rijke, M. (2005). *Query formulation for answer processing*. Published research, Informatics Institute, University of Amsterdam. Retrieved from http://dare.uva.nl

Resnik, Philip. (1995). Disambiguating Noun Groupings with Respect to WordNet Senses. *Third Workshop on Very Large Corpora*. Retrieved from http://acl.ldc.upenn.edu/W/W95/W95-0105.pdf