# NLP for Privacy Policy Document Analysis

Aaron Massey, UMBC

Alden Dima, NIST

# NLP in Requirements Engineering

## Mid-2010s: NLP is Resurgent

"[W]e witness the novel golden age of NLP technologies [...] **It is therefore an appropriate moment to create a venue in which researchers on applications of NLP to RE problems can meet**, share ideas and create synergies, assisted by experts from the NLP community."

F. Dalpiaz, A. Ferrari, X. Franch, and C. Palomares, Call for Papers, 1st Workshop on Natural Language Processing for Requirements Engineering, 2018

## Mid-1990s: NLP is Worthless

"[N]atural language does not now, nor will it in the foreseeable future, provide a level of understanding that could be relied upon, and even if it could, **it is highly questionable that the resulting system would be of great use in requirements engineering.**"

K. Ryan, "The Role of Natural Language in Requirements Engineering," in Proceedings of the IEEE International Symposium on Requirements Engineering, 1993, pp. 240–242.

# Primary Ongoing Challenge for NLP in SE/RE

"[P]erhaps a dumb tool doing an identifiable part of such a task may be better than an intelligent tool trying but failing in unidentifiable ways to do the entire task."

D. Berry, R. Gacitua, P. Sawyer, and S. F. Tjong, "The Case for Dumb Requirements Engineering Tools," in *Requirements Engineering: Foundation for Software Quality*, B. Regnell and D. Damian, Eds. Springer Berlin Heidelberg, 2012, pp. 211–217.

**Measurement of NLP techniques are misleading** for Software Engineering and Requirements Engineering:

**Option 1:** Break the problem down into smaller tasks with a smaller penalty

**Option 2:** Adjust measurement techniques to account for SE/RE

**Option 3:** Focus on Human-in-the-Loop systems

# Automated Text Mining for Requirements Analysis of Policy Documents

**Goal: Identification of documents relevant to specific requirements**

Can automated text mining help requirements engineers determine whether a policy document contains requirements expressed as either privacy protections and vulnerabilities?

Uses a large corpus of privacy policy documents:

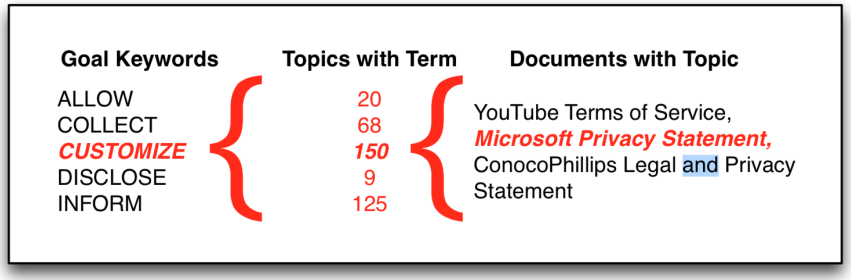
> The RE 2013 Policy Document Corpus consists of 2,061 Privacy Policies, Terms of Use, Terms and Conditions, and Terms of Service documents, and other policy documents

*Aaron K. Massey, Jacob Eisenstein, Annie I. Antón, and Peter P. Swire. "Automated Text Mining for Requirements Analysis of Policy Documents" 21st IEEE International Requirements Engineering Conference. Rio de Janeiro, Brazil, July 2013.*

# Automated Text Mining for Requirements Analysis of Policy Documents (2)

**Methodology**

1. Created a Latent Dirichlet Allocation (LDA) model with k = 154 topics using R topicmodels
    a. Split docs: training (90%) and holdout (10%)
    b. Generated 35 topic models with varying k and picked one with lowest perplexity
2. Identified documents addressing concerns using goals-based RE




NUMBER OF POLICY DOCUMENTS (OUT OF 2,061) IDENTIFIED AS POTENTIALLY CONTAINING GOAL STATEMENTS

| Key-word | Docu-ments | Key-word | Docu-ments | Key-word | Docu-ments |
|---|---|---|---|---|---|
| access | 904 | apply | 331 | change | 31 |
| collect | 202 | comply | 339 | connect | 121 |
| display | 308 | help | 61 | honor | 19 |
| inform | 23 | limit | 52 | notify | 347 |
| opt-in | 32 | opt-out | 76 | post | 76 |
| request | 31 | reserve | 51 | share | 300 |
| specify | 38 | store | 38 | use | 525 |

**Can limit search from 2,061 documents to less than 100 related to given goal keyword**

# Analysis of Privacy Concerns Across Industry

Researchers and regulators: **Can we identify privacy concerns within policy documents at scale?**

Manual analysis of privacy policy collections

- Example: Professor and eight law students reviewed **249** policies (Marotta-Wurgler, 2016)

Is automated replication possible?

- RE 2013 Policy Document Corpus has 2061 documents (Massey et al, 2013)
- Web crawling enables large collections
  - ✦ 1M policy documents

**Challenges**

**Keyword Search→False Positives**

**Want to quickly re-analyze**

- New concerns
- New examples

**Classifiers**

- Imbalanced classes
- Many concerns→many classifiers
- Lack of labelled training data
  - ✦ Often small sets of examples
  - ✦ Deep Learning→ Large training sets + time

# Methodology

1. Extract sentences from training data
2. Create Latent Semantic Indexing (LSI) model of sentences
3. Filter training data using keywords
4. Summarize filtered data with LexRank
5. Manually select exemplar (seed) sentences from filtered data
6. Search of new corpus using exemplar sentences and LSI model
7. Select sentences above cosine similarity threshold

**We can identify relevant individual sentences from a collection of documents!**

# Summary

NLP can play an important role in Requirements Engineering, but the benefits and challenges of this approach remain mostly a topic for academic research.

# Thank you!

**Aaron Massey**

Assistant Professor of Software Engineering

Information Systems Department

University of Maryland Baltimore County

**akmassey@umbc.edu**

https://userpages.umbc.edu/~akmassey/