Leg Text Scraper Technology Review

Katherine Chang, Shujie Chen, Gan Gao, Cynthia Wu
Fall 2021, CSE 583

Background

Goals:

- Build a tool to scrapes audio files from state legislature committee hearing and prepares its for text analytics
- Build a dashboard of text analytics to indicate the power of this text corpora and to provide high-level consumable information about a topic of interest

Use Cases:

- Researchers (Raw text data from audio)
- Members of the Public (Data dashboard focused on a specific topic)

Technology: Speech-to-Text

- **Google Speech-to-Text**
  - Closed source, cloud-based (requires Google Cloud account)
  - Price: $0.006 / 15 seconds
  - Includes speaker diarization
  - Generated several example transcripts to compare to open-source options
- **DeepSpeech**
  - Open-source, off-line, on-device,
  - Based on Baidu Research DeepSpeech system
  - Pros: Pre-trained speech recognition models, flexibility to train own models
  - Limitations: No speaker diarization

Similarity between Open-Source and Closed-Source

- spaCy's pre-trained English pipelines helps gauge the similarity of between our Google generated speech to audio transcripts and the DeepSpeech generated transcripts.
- Results reply on pretrained pipeline and model used:

	Deep Speech	Google Speech-to-Text
Small context-sensitive tensor approach	78.5%	
Word vectors similarity approach	98.9%	

Technology: NLP/Text Analytics

- nltk
- spaCy
- Gensim
- coreNLP
- TextBlob
- Pattern

Technology: NLP/Text Analytics

- **nltk**

  Pros: most well-known; full NLP libraries with many 3rd extensions; supports many languages

  Cons: Slow; Only splits text by sentences without analyzing semantics

- **spaCy**

  Pros: Fast; Includes word-tokenization

  Cons: Less flexibility

Choice

- Webscraping: Selenium
  - Most popular
- Speech-to-Text: DeepSpeech
  - Open-source, off-line, on-device
- Text Analysis: nltk
  - Most popular, ease of implementation, plenty of resources
  - Interactive Visualizations via Altair