NATURAL LANGUAGE PROCESSING

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

NLP and Language Models

What is this?

Prerecorded videos + exercises you can integrate into your class

https://www.cs.utexas.edu/~gdurrett/courses/nlp-module/

What is NLP?

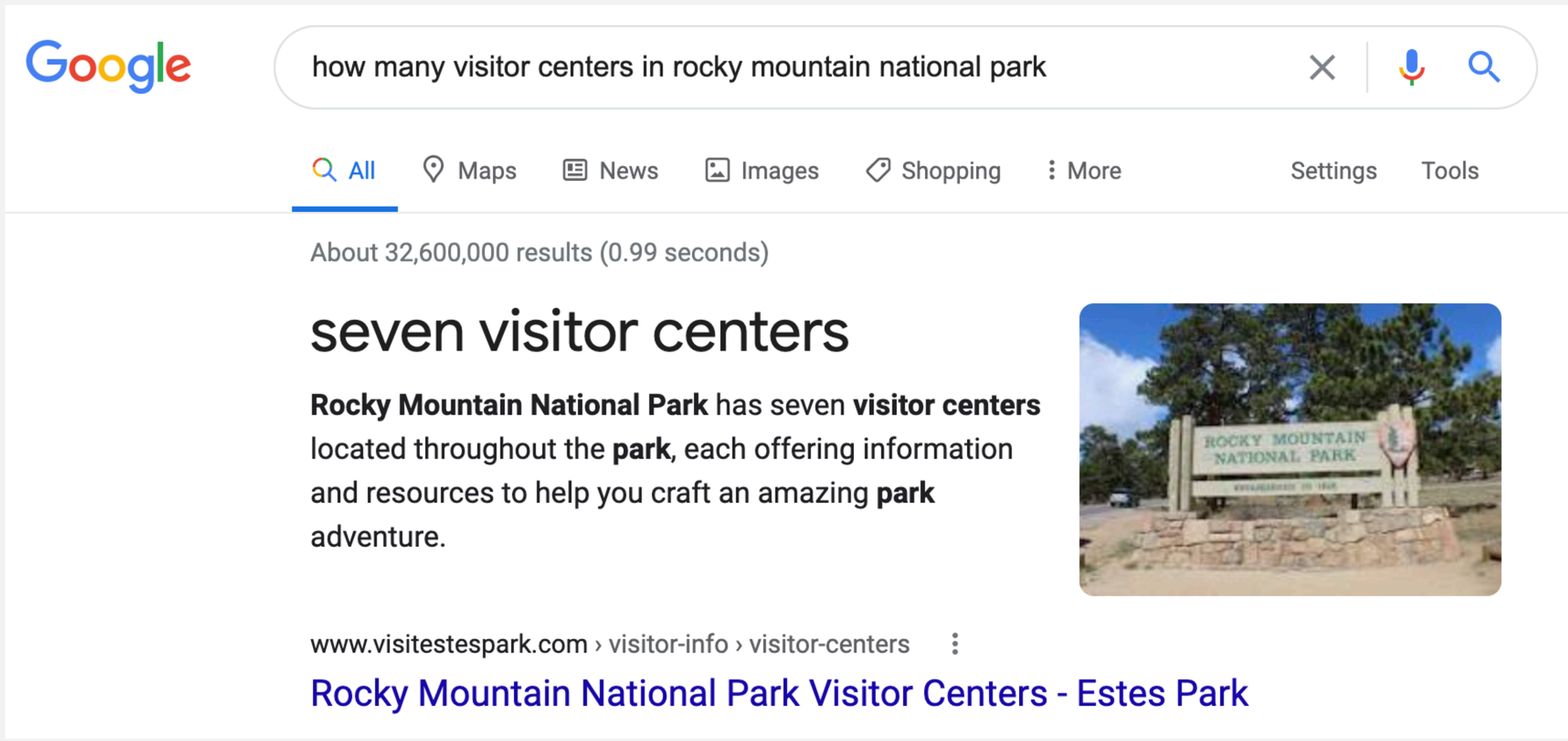
Natural Language	Processing
human languages (not computer languages)	doing things with them automatically!

Everyday NLP Tools

- Google Translate

- Siri/Alexa/etc.

- Even some Google searches are powered by NLP:



What can we do with NLP?

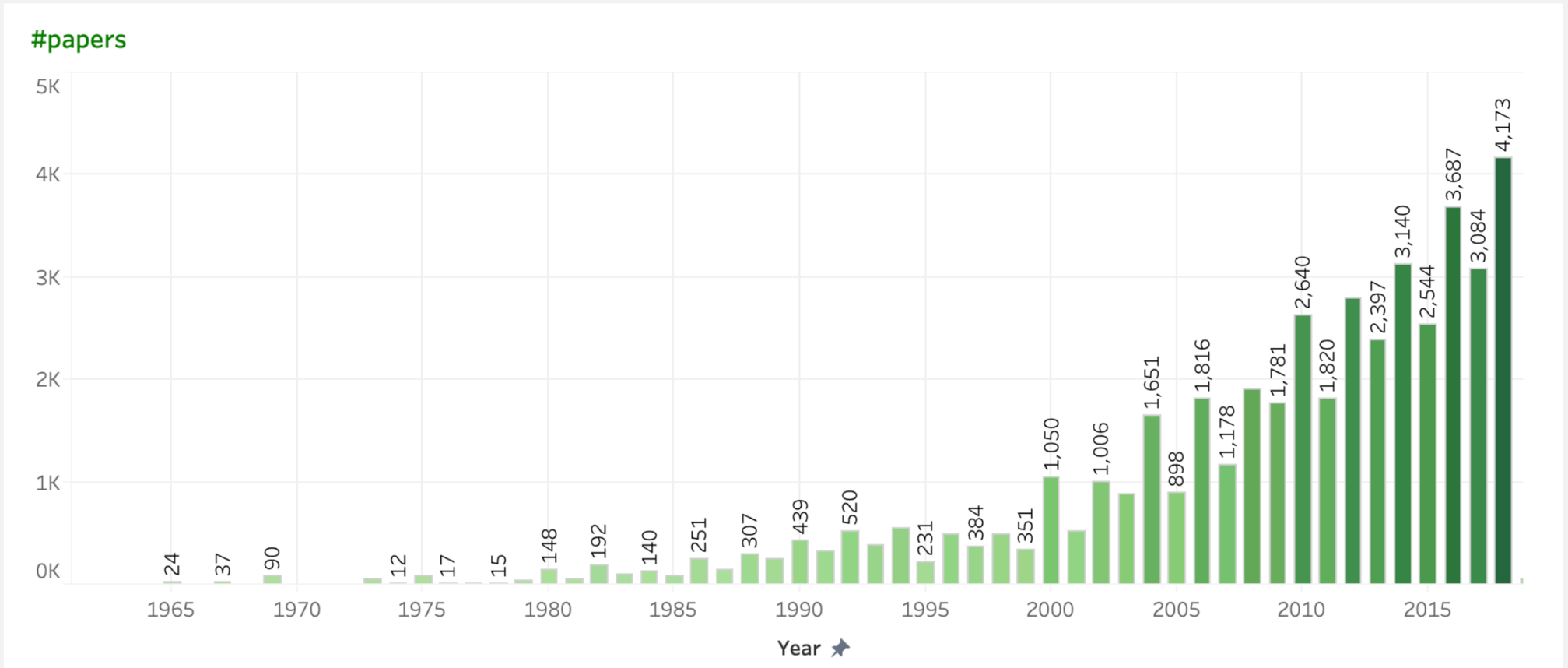
- Data science: all kinds of language data to analyze!
  - Product reviews
  - Customer service chat logs
  - Social media: emerging information about events, spread of information, …

- It’s easier than ever to do it: you no longer need an MS/PhD to build some basic NLP models
  - Sentiment Analysis tutorial: `https://realpython.com/sentiment-analysis-python/`

- Important for students: be aware of how to look at data, know about NLP systems and what they can do

Why NLP?

- Many colleges offering NLP now, huge hiring boom in professors
- Many more MS/PhD students studying it (and still getting jobs, for now)
- Growth in published papers:



Mohammad (2019)

What I'm Offering

- 10 videos / 80 minutes of prerecorded video

- Two activities:
  - Building an *n-gram language model*: a probabilistic model for "predictive text" (what's the next word after I've typed a few?). Only relies on data structures and basic probability calculations [30-90 minute coding exercise]
    - **Conditional probability distribution:** P(next word = *y* | previous word = *x*)

      P(next word = *Austin* | previous word = *to*) = 0.2

      "if we see *to* I think there's a 20% chance the next word is *Austin*"
  - Play around with Write With Transformer [15 minute web demo exercise]
    - Large neural network model that's really good at this task w/a slick demo

https://www.cs.utexas.edu/~gdurrett/courses/nlp-module/

Goals of this module

- Learn what NLP is about
- Learn some basic ideas of machine learning: fitting a statistical model to examples of a problem we want to solve to learn how to solve that problem
- See how a statistical model for predictive text works (what word should come next in this sentence?)
- Learn the connections between this language model and state-of-the-art systems such as Google's BERT and OpenAI's GPT-3 models

https://www.cs.utexas.edu/~gdurrett/courses/nlp-module/

Demo/Walkthrough

https://www.cs.utexas.edu/~gdurrett/courses/nlp-module/

Version 1 (60-70 minutes)

- Intro and a high-level sketch of how some source code works [~35 minutes]
- *Skip the coding exercise*
- Play around with Write With Transformer [15 minute web demo exercise]
- Discussion/conclusion [15 minutes]
- We ran this at the UT Summer Academies for an audience of ~30

https://www.cs.utexas.edu/~gdurrett/courses/nlp-module/

Version 2 (120 minutes)

- Intro [~20 minutes]
- Prep for coding [20 minutes]
- Coding [30-90 minutes, can be asynchronous]
- Play around with Write With Transformer [15 minute web demo exercise]
- Discussion/conclusion [15 minutes]

https://www.cs.utexas.edu/~gdurrett/courses/nlp-module/

Thanks

- Marla Gilliland, Colin Pittman, and other folks at UT LAITS

IFML Institute for Foundations of MACHINE LEARNING

- Verena Kalhoff and Adam Klivans (IFML)

- Joy Schwartz (TACC)

- Jose Luis Silva-Smith (Tornillo High School)

Email me if you want to use this!

https://www.cs.utexas.edu/~gdurrett/courses/nlp-module/