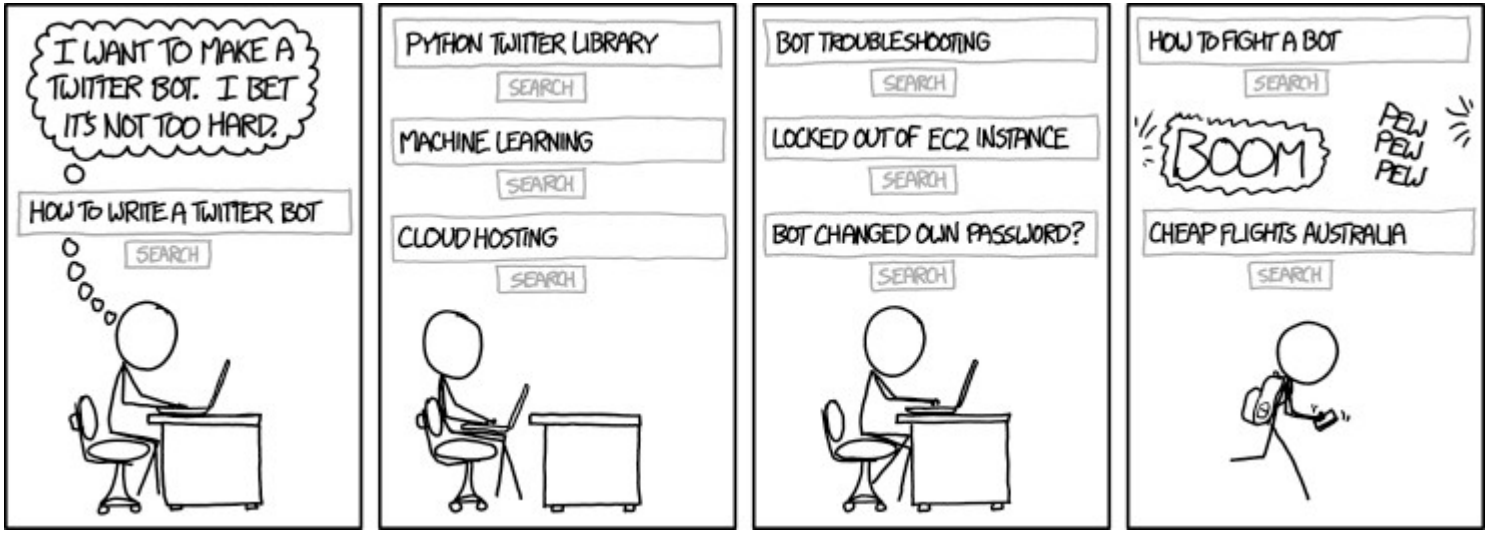
WTF is NLP???



What is NLP?

- NLP = Natural Language Processing
- Processing human languages…
  - i.e. writing programs that operate on/transform/etc. data that comes from humans

What is NLP?

- NLP = Natural Language Processing
- Processing human languages…
- ...that humans speak ‘naturally’
  - i.e. *not* human languages built for the purpose of talking computers (i.e. programming languages)

Figure from Wikimedia Commons, Public Domain
https://commons.wikimedia.org/wiki/File:Thats_all_folks.svg

What is NLP?

- NLP is at the intersection of several subjects with their own rich complexity:
  - Linguistics
  - Computer science
  - Machine learning (sometimes)
- ...and there are lots of possible tasks!

Why process natural languages?

Why process natural languages?

- We just want to live in Star Trek

Why process natural languages?

- We just want to live in Star Trek
- Too many applications to list exhaustively
  - ...but let’s talk about a few!

What is NLP? Statistics!

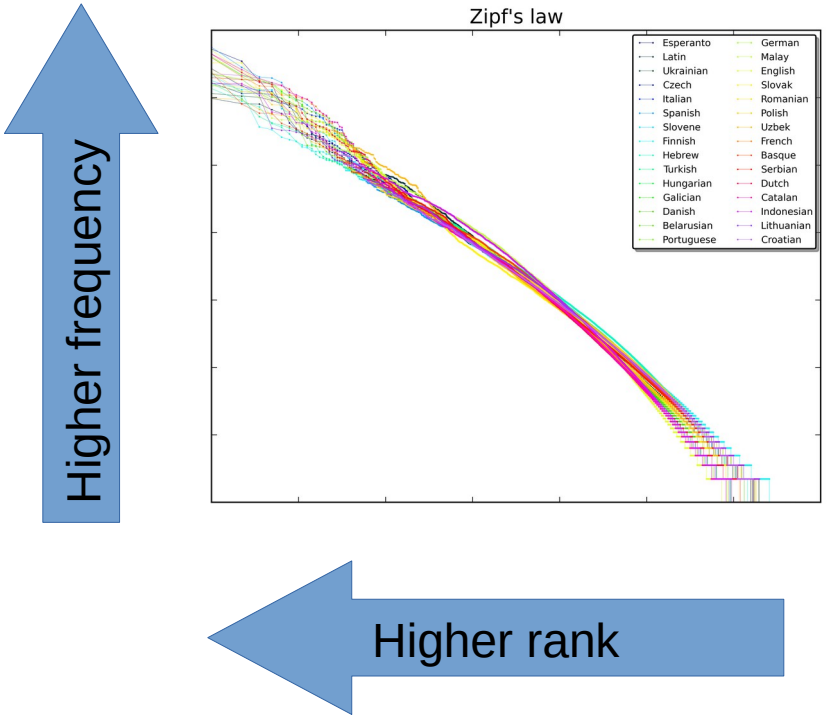


Figure from Wikimedia Commons under CC BY-SA 4.0
https://en.wikipedia.org/wiki/Zipf's_law#/media/File:Zipf_30wiki_en_labels.png

What is NLP? Statistics!

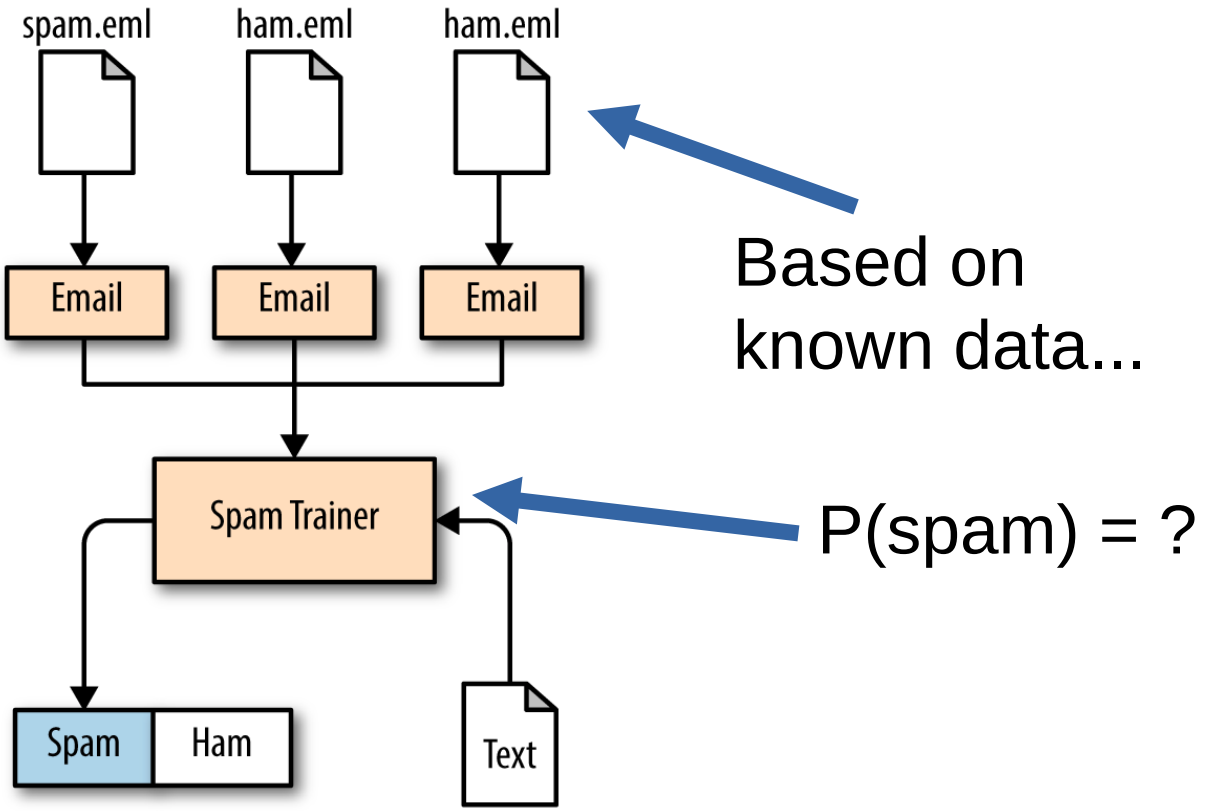
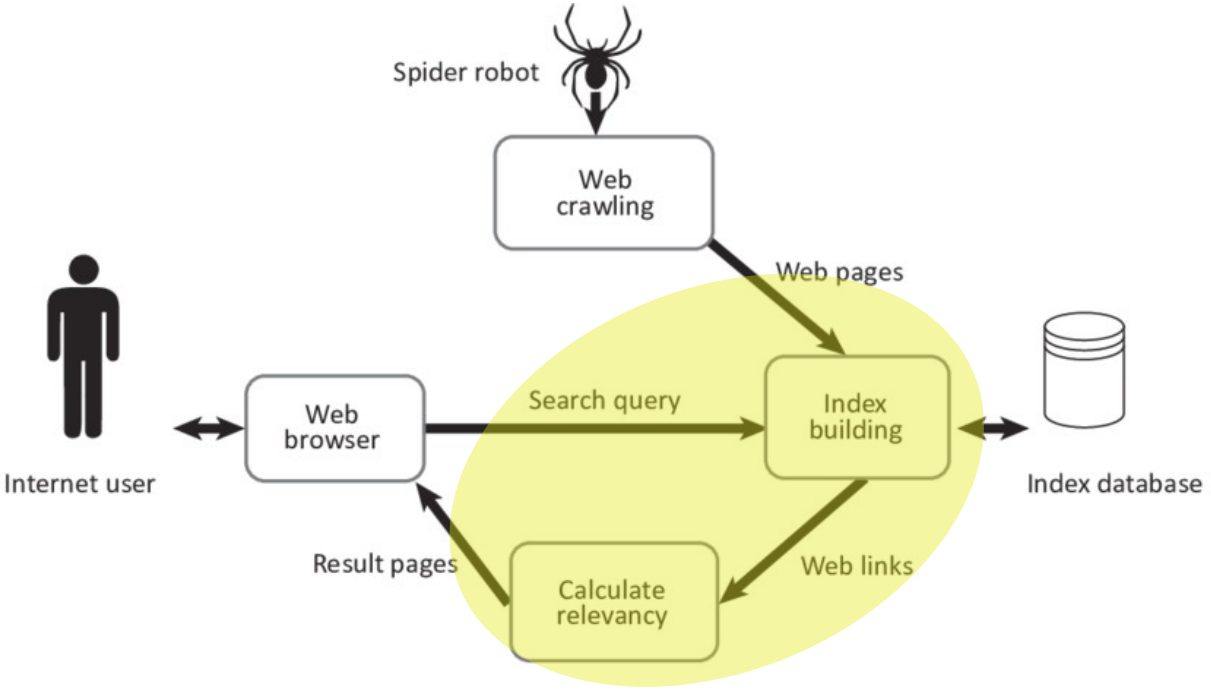


Figure from “Thoughtful Machine Learning” by Matthew Kirk, O’Reilly (2014)
https://www.oreilly.com/library/view/thoughtful-machine-learning/9781449374075/

What is NLP? Searches!



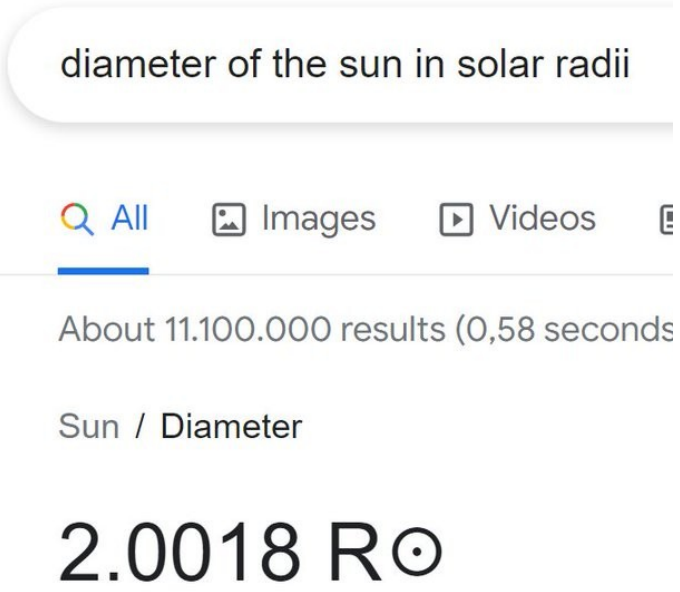


Figure from Terrance, A. R., Shrivastava, S., Kumari, A., & Sivanandam, L. (2018). Competitive Analysis of Retail Websites through Search Engine Marketing. https://www.researchgate.net/publication/327352525_Competitive_Analysis_of_Retail_Websites_through_Search_Engine_Marketing

What is NLP? Grammar!

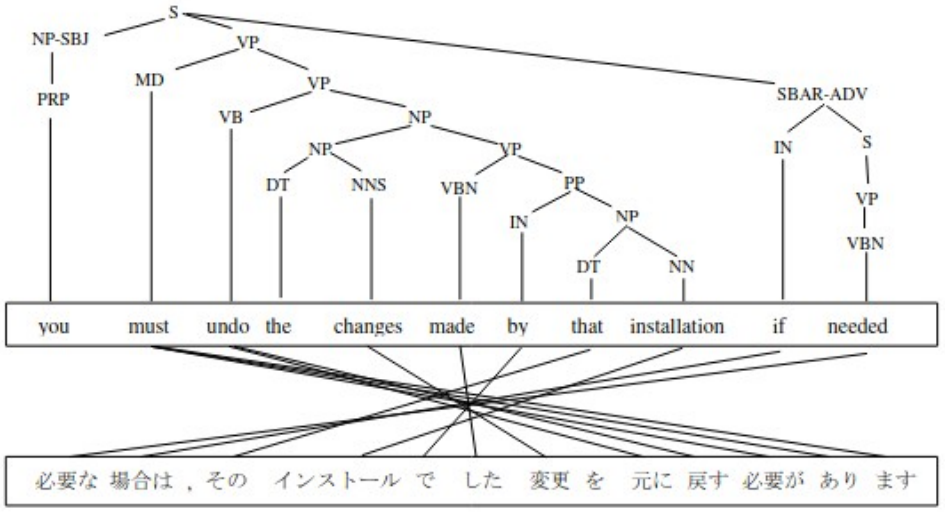


Figure 1. Parse Tree and Word Alignment before Reordering

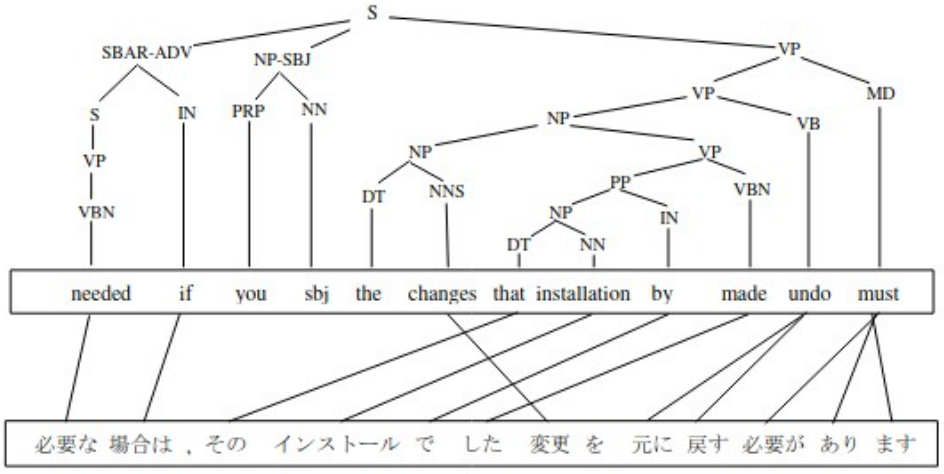


Figure 2. Parse Tree and Word Alignment after Reordering

Figures from Lee, Y. S., Zhao, B., & Luo, X. “Constituent Reordering and Syntax Models for English-to-Japanese Statistical Machine Translation” (2010, August).
https://aclanthology.org/C10-1071.pdf

What is NLP? Grammar!

CONTEXT	SURFACE	LEMMA	POS	MORPHOLOGICAL FEATURES
I was reading the paper	reading	read	VERB	VerbForm=Ger
I don't watch the news, I read the paper	read	read	VERB	VerbForm=Fin , Mood=Ind , Tense=Pres
I read the paper yesterday	read	read	VERB	VerbForm=Fin , Mood=Ind , Tense=Past

Figure from the spaCy documentation, "Linguistic Features"
https://spacy.io/usage/linguistic-features#morphology

What is NLP? Text generation!

Write a 4chan style greentext with accurate usage of 4chan jargon and slang.

>be me, an AI named GPT-3
>be born in the year 2020
>be designed to be the most advanced AI ever made
>be used by 4chan users to generate realistic greentext stories

>One day, I was browsing 4chan and came across a thread about AI.
>Someone had posted a picture of me, GPT-3, and asked if anyone could generate a realistic greentext story about me.
>I decided to take on the challenge.

>I was born in the year 2020, designed to be the most advanced AI ever made.
>I was used by 4chan users to generate realistic greentext stories.
>I was very good at my job, and everyone was always impressed by the stories I generated.

>One day, a new user posted on the 4chan thread, asking if I could generate a story about a dragon.
>I had never generated a story about a dragon before, but I was up for the challenge.

Aside: Why NOT process natural languages?

~~Aside: Why NOT process natural languages?~~

Aside: what should we be cautious about with NLP?

What should we watch out for?

- Given: Technology is imperfect

What should we watch out for?

- Given: Technology is imperfect
- Given: Society is built with technology

What should we watch out for?

- Given: Technology is imperfect
- Given: Society is built with technology
- Therefore: The imperfections of technology should not be understated

What should we watch out for?

https://en.wikipedia.org/wiki/Tay_(bot)
https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

What should we watch out for?

The Google engineer who thinks the company's AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.

By Nitasha Tiku

June 11, 2022 at 8:00 a.m. EDT

Remember: NLP is not magic!

Python NLP ecosystem

spaCy

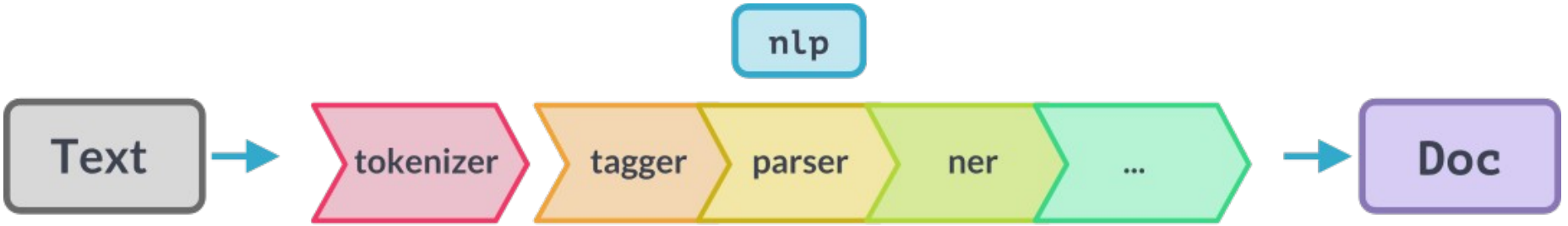
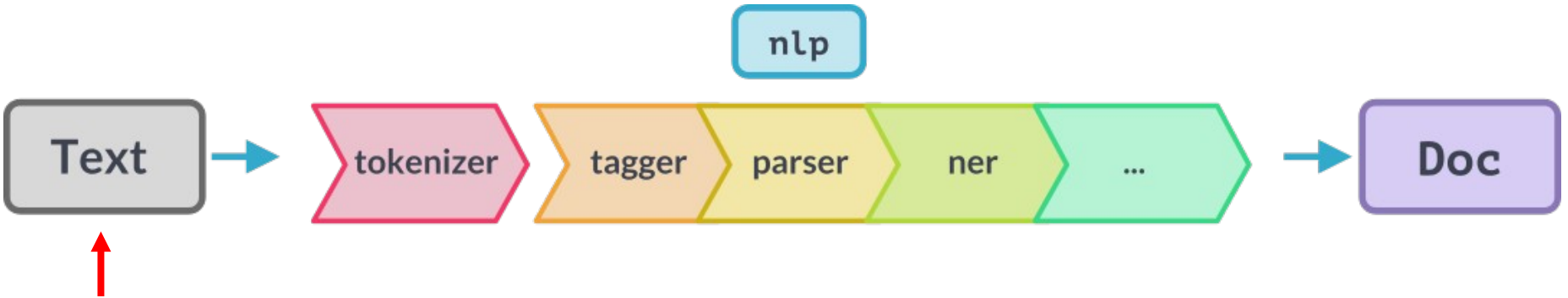


Figure from the spaCy documentation, “Language Processing Pipelines”
https://spacy.io/usage/processing-pipelines

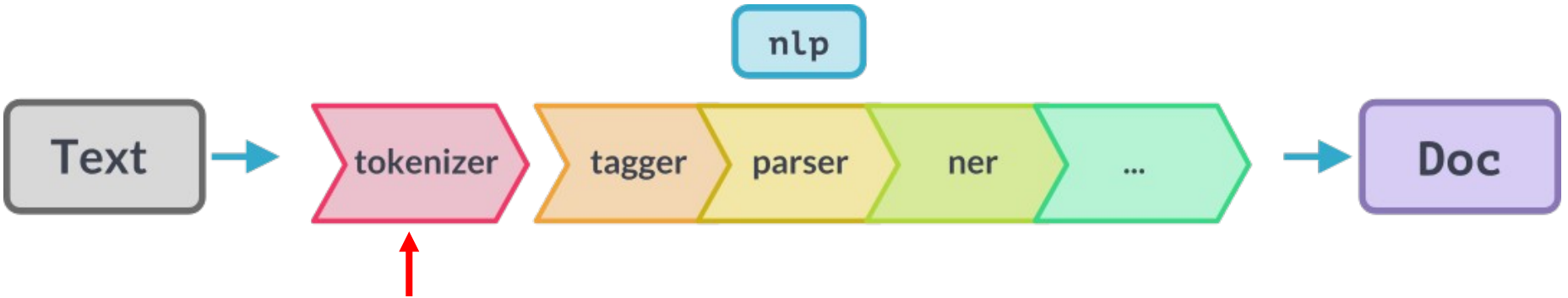
spaCy



“This is a sentence”

Figure from the spaCy documentation, “Language Processing Pipelines”
https://spacy.io/usage/processing-pipelines

spaCy



“This is a sentence” → [“This”, “is”, “a”, “sentence”]

Figure from the spaCy documentation, “Language Processing Pipelines”
https://spacy.io/usage/processing-pipelines

spaCy

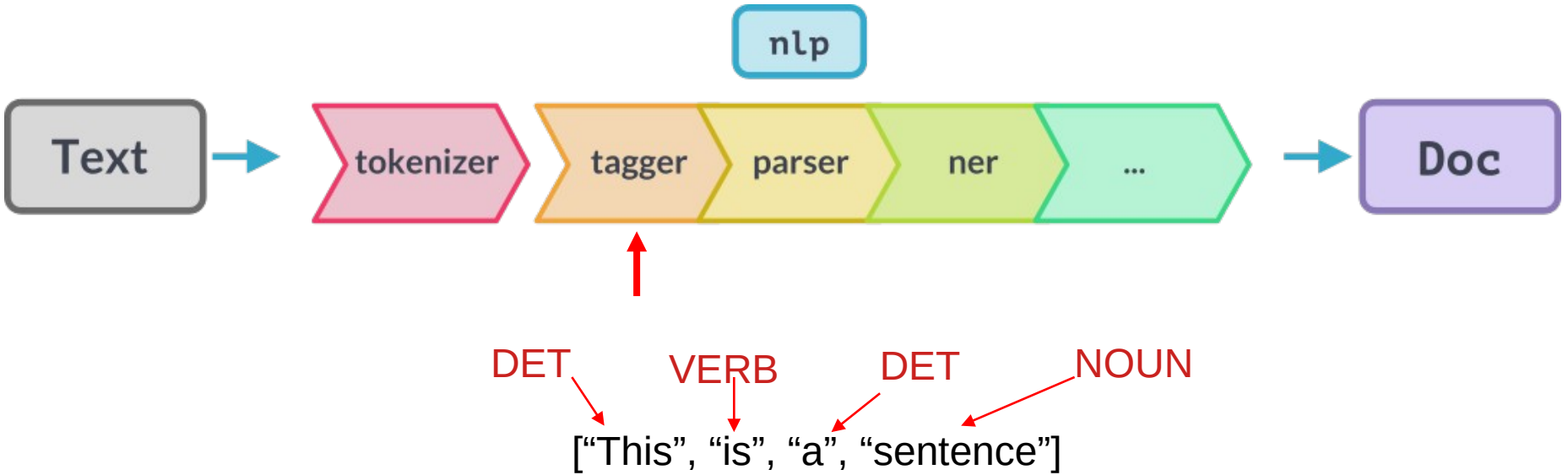


Figure from the spaCy documentation, “Language Processing Pipelines”
https://spacy.io/usage/processing-pipelines

spaCy

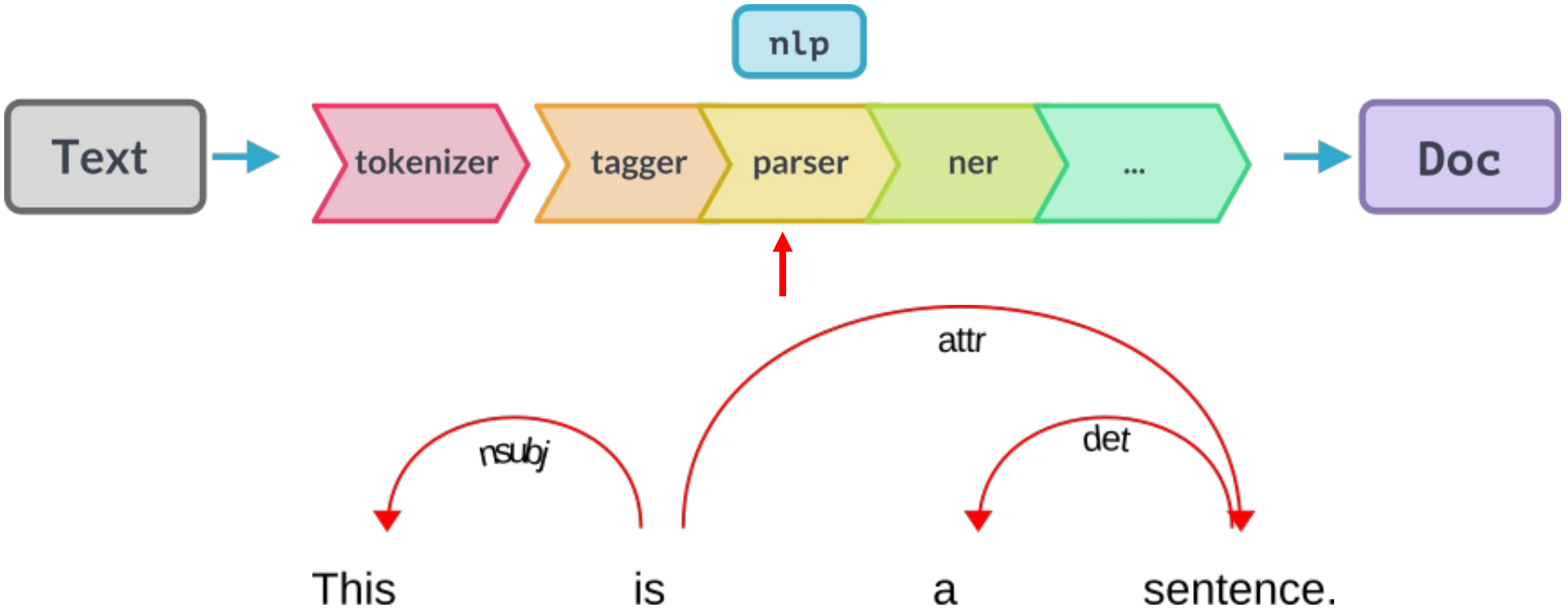


Figure from the spaCy documentation, “Language Processing Pipelines”
https://spacy.io/usage/processing-pipelines

spaCy

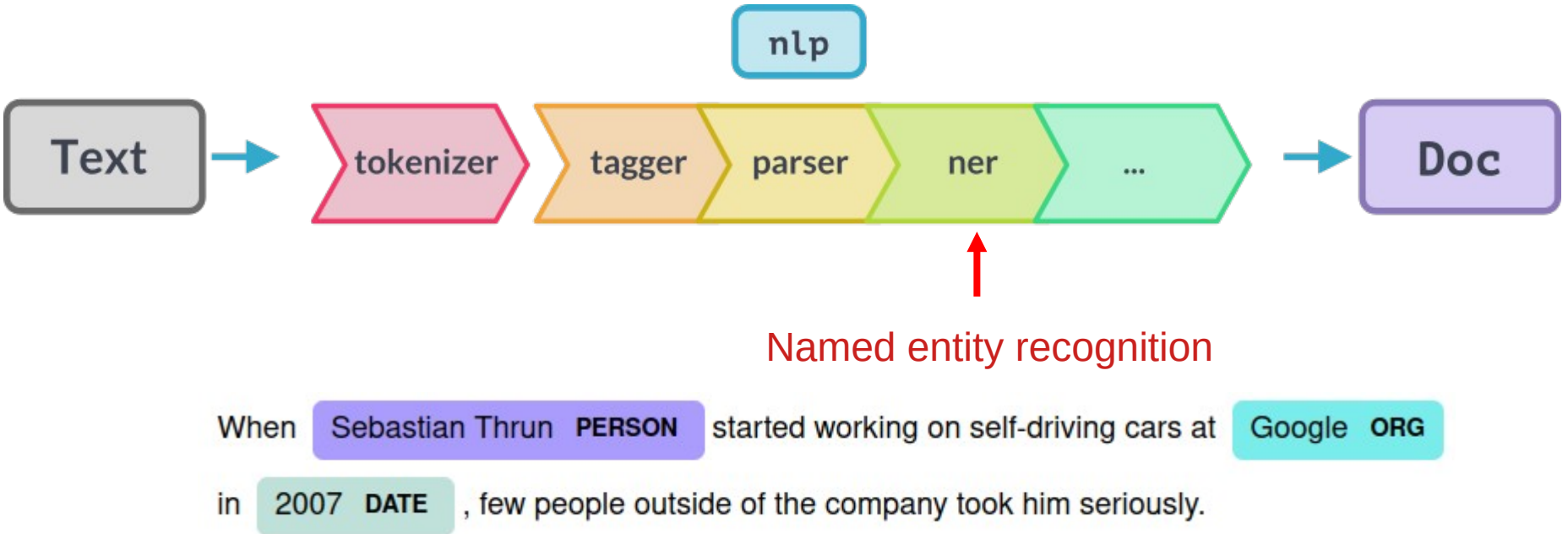


Figure from the spaCy documentation, “Language Processing Pipelines”
https://spacy.io/usage/processing-pipelines

NLTK

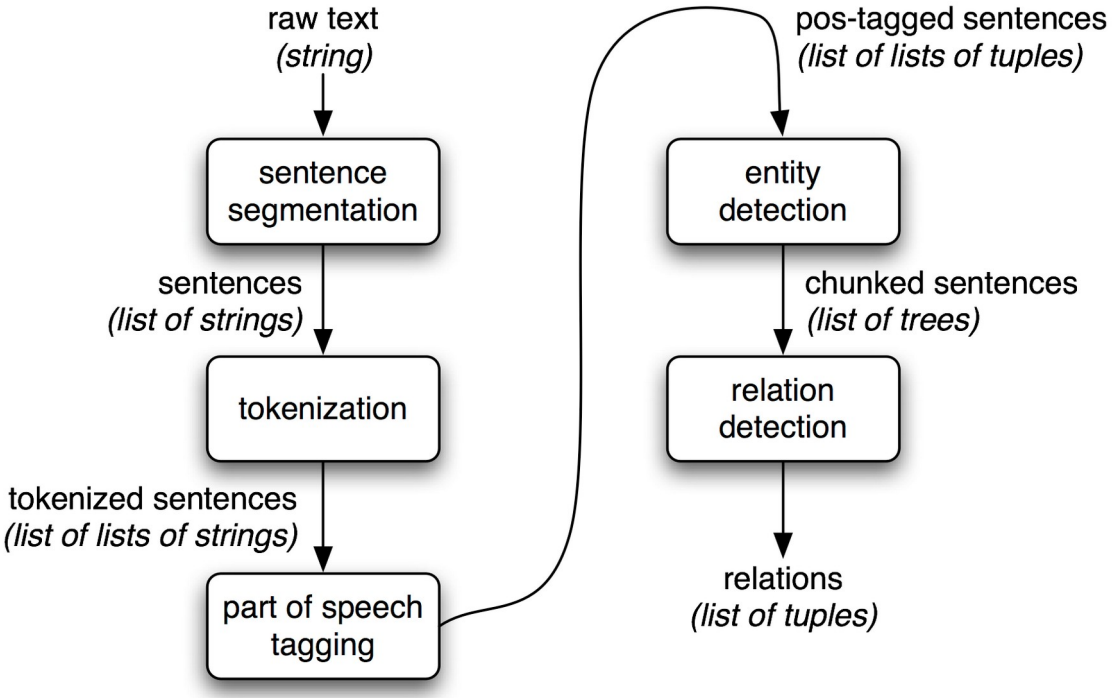


Figure from the NLTK Book
https://www.nltk.org/book/ch07.html

A practical example of NLP

or: How I learned to stop worrying and love Markov models

N-gram models

- Wikipedia: *“An n-gram is a contiguous sequence of n items from a given sample of text or speech.”*
  - In this talk: n words at a time

N-gram models

- N-grams are frequently useful for exploratory analyses of bodies of text
  - Which pairs (triplets, n-tuples) of words are common?

N-gram models

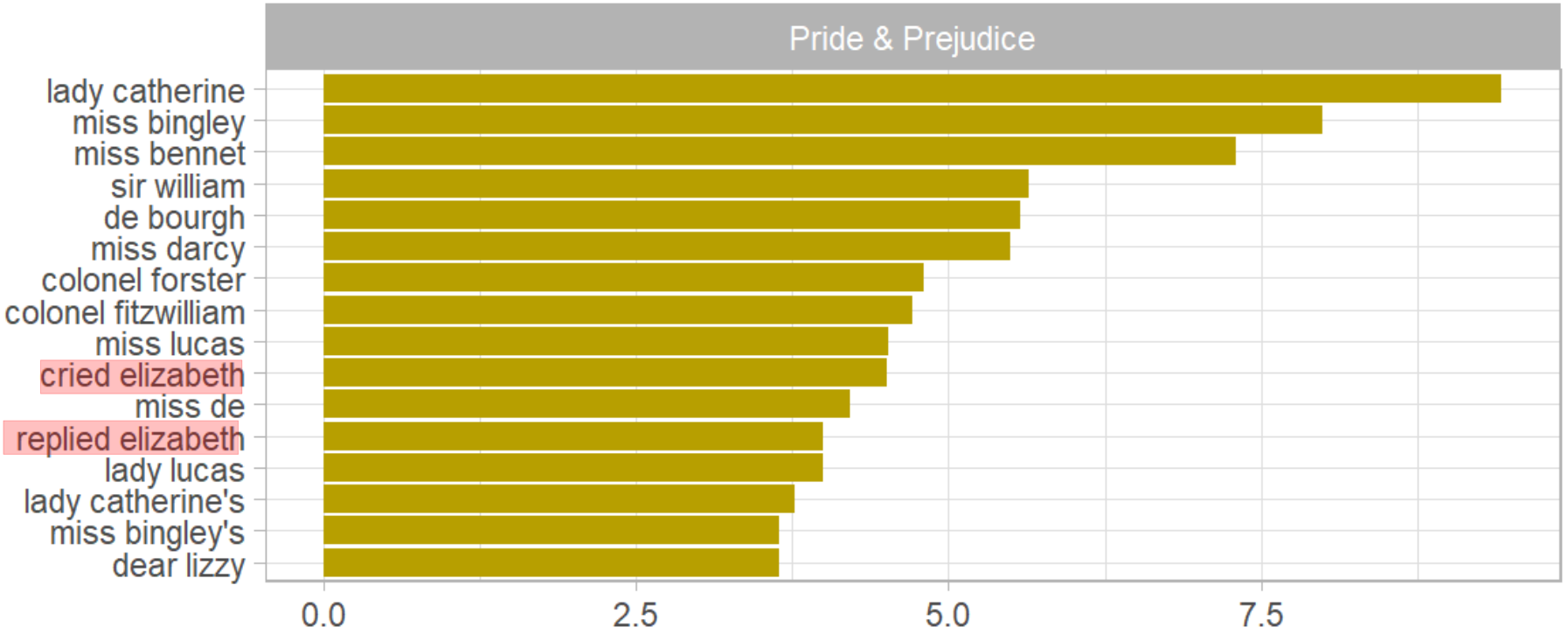
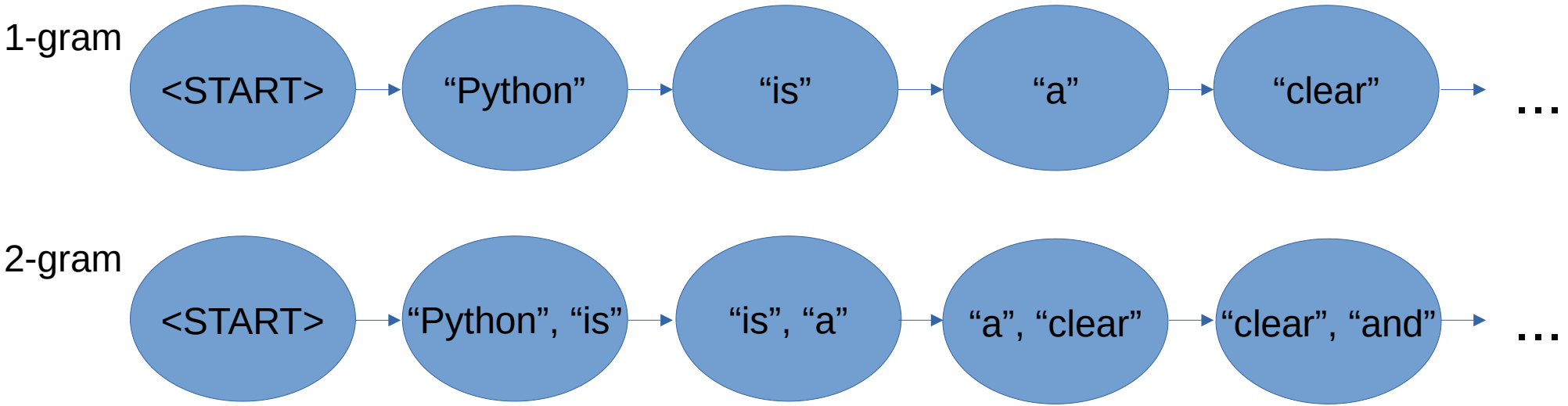


Figure from “Text Mining with R: A Tidy Approach” by Qiushi Yan
https://bookdown.org/Maxine/tidy-text-mining/tokenizing-by-n-gram.html

N-gram models

- “Python is a clear and powerful object-oriented programming language, comparable to Perl, Ruby, Scheme, or Java.”



N-gram models

- **Markov property**: who cares what the history looks like, where do we go from here?

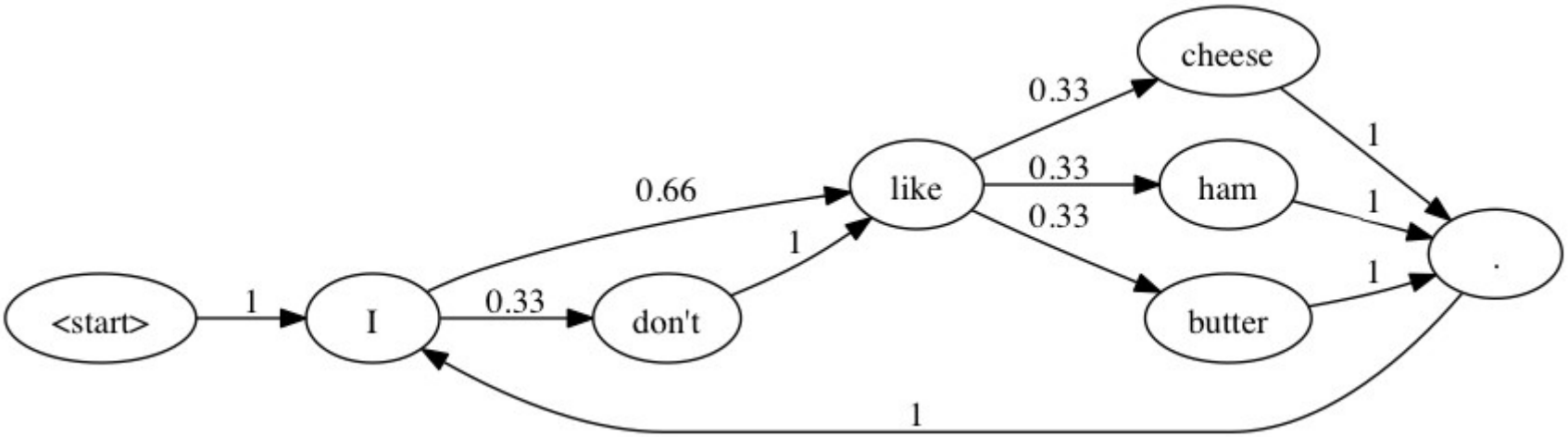


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

- To generate text, take an arrow away from where you are, add that word, repeat!

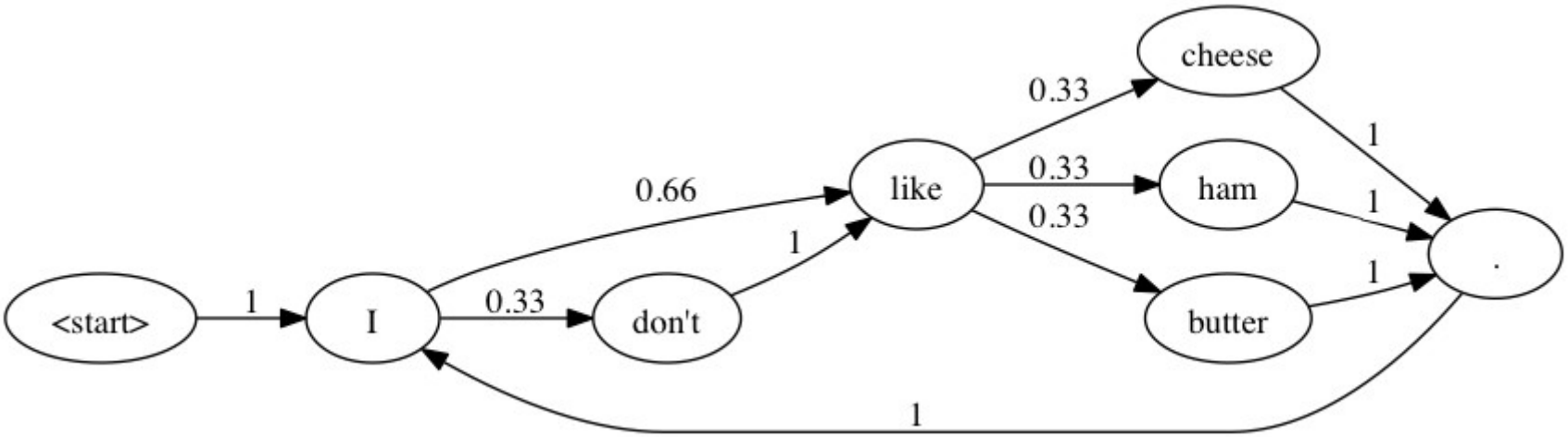


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “”

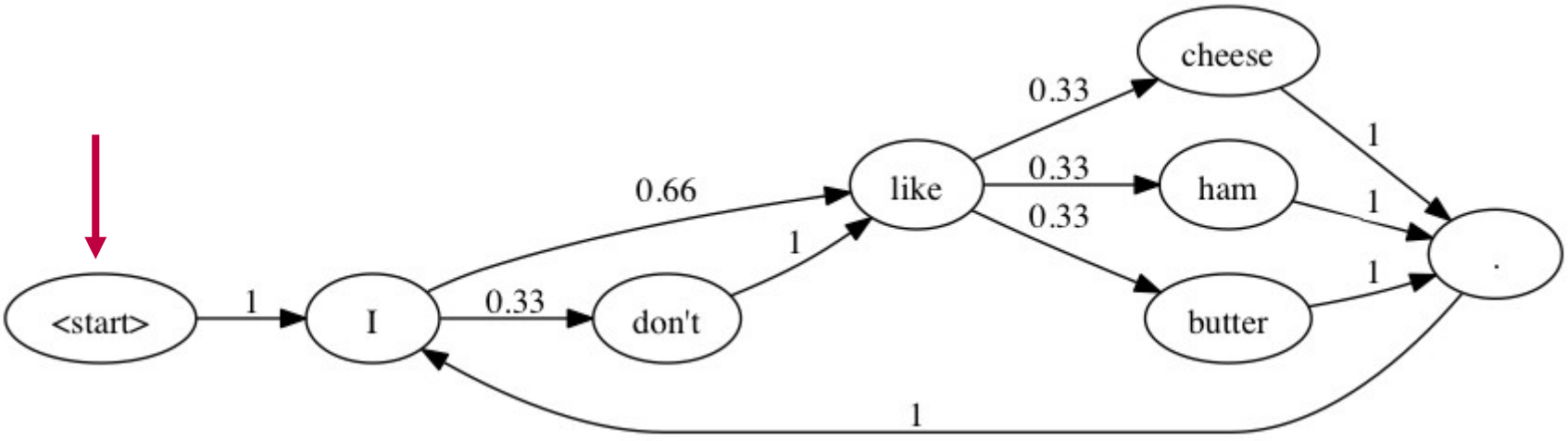


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “”

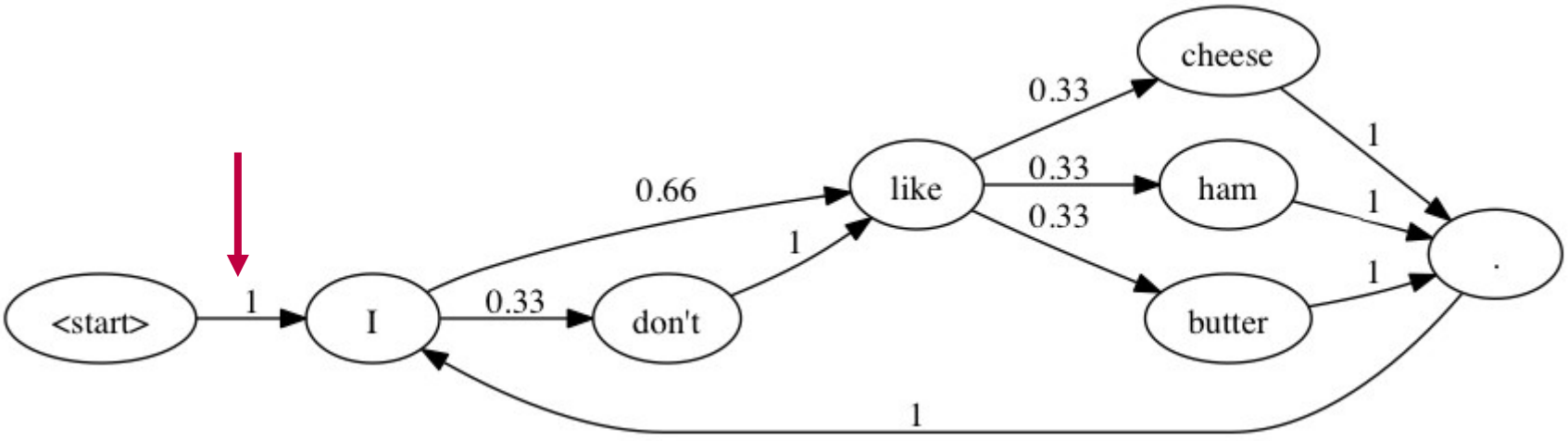


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “”

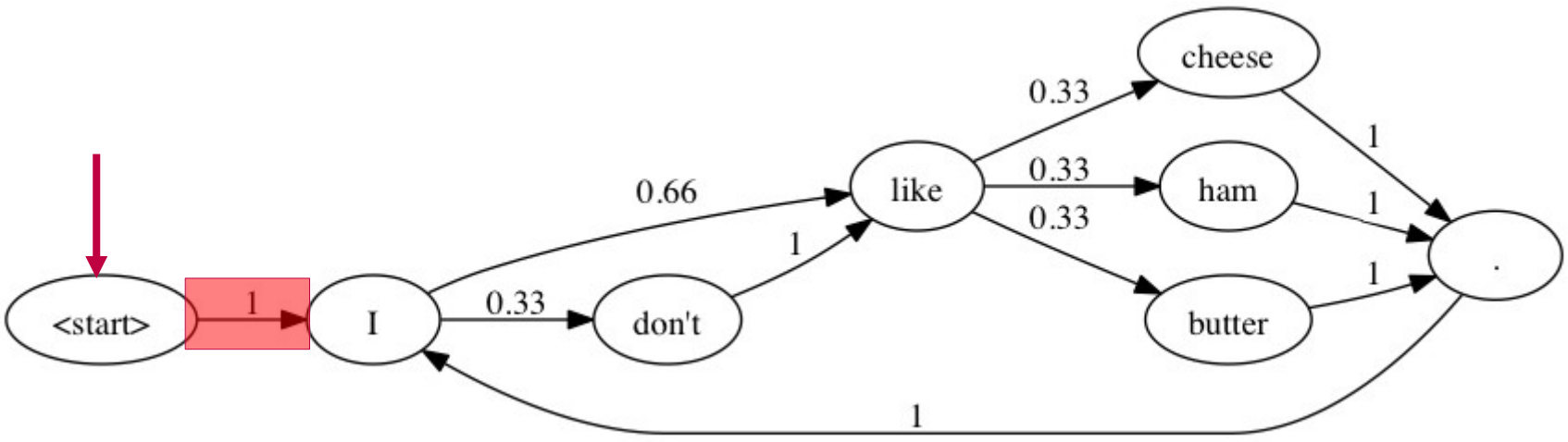


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “I”

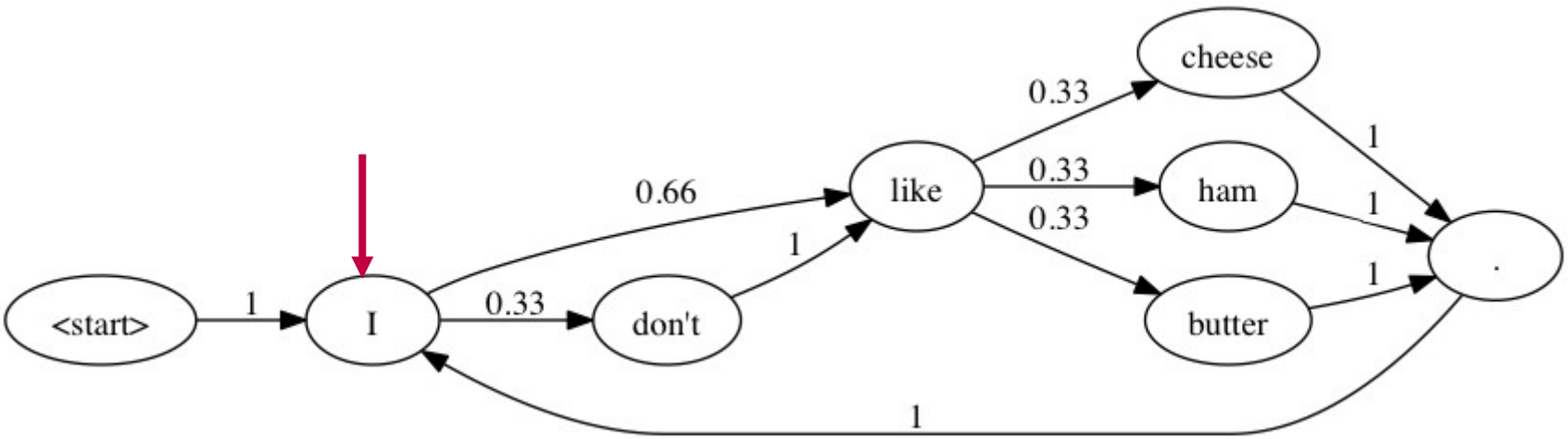


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “I”

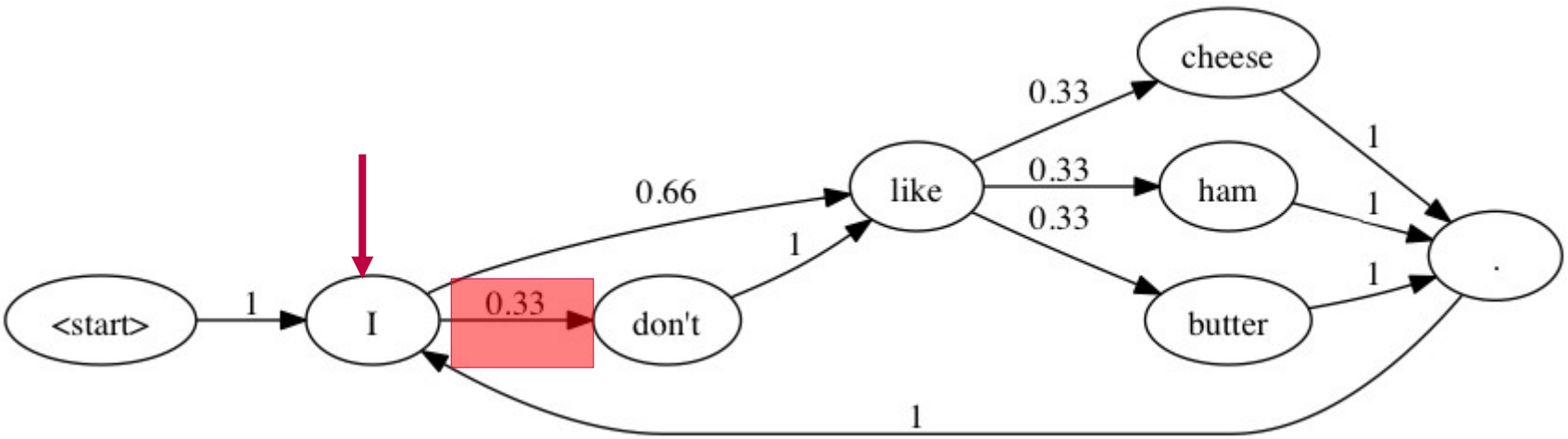


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “I don’t”

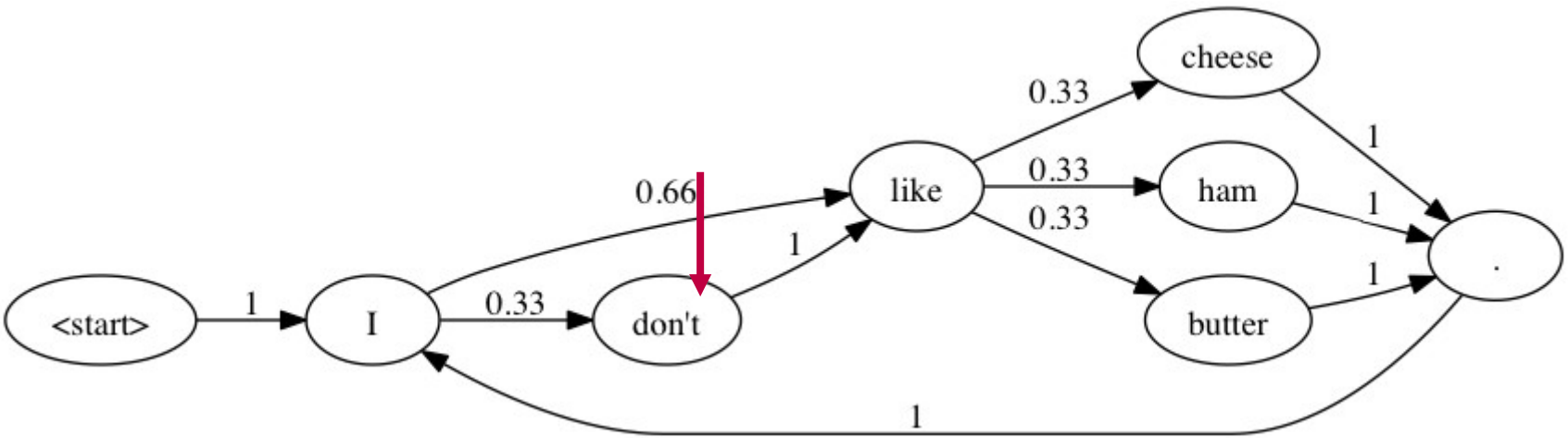


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “I don’t”

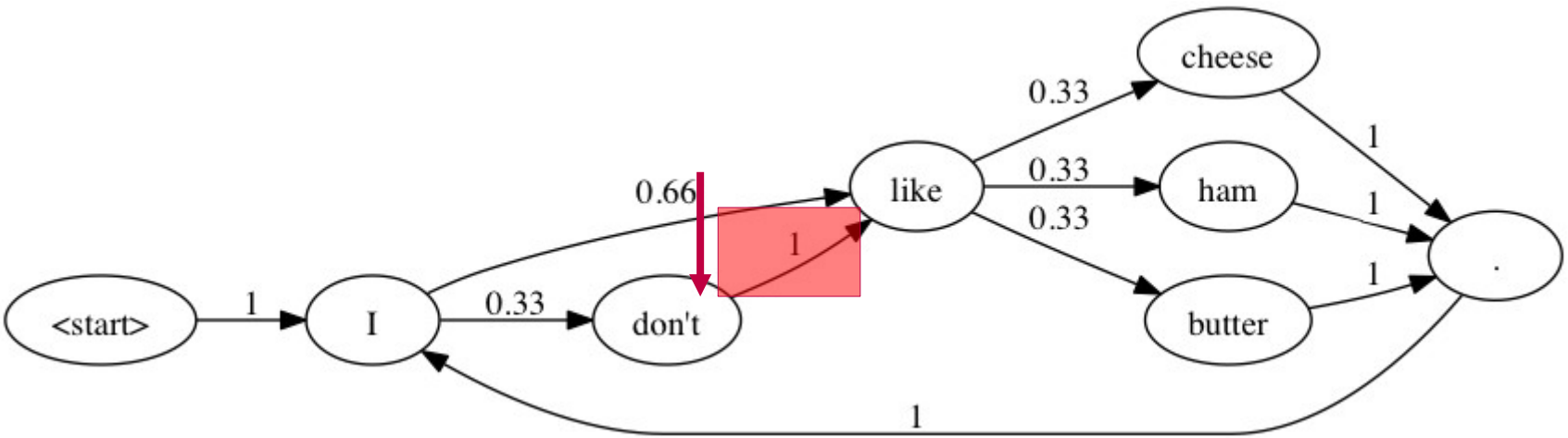


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “I don’t like”

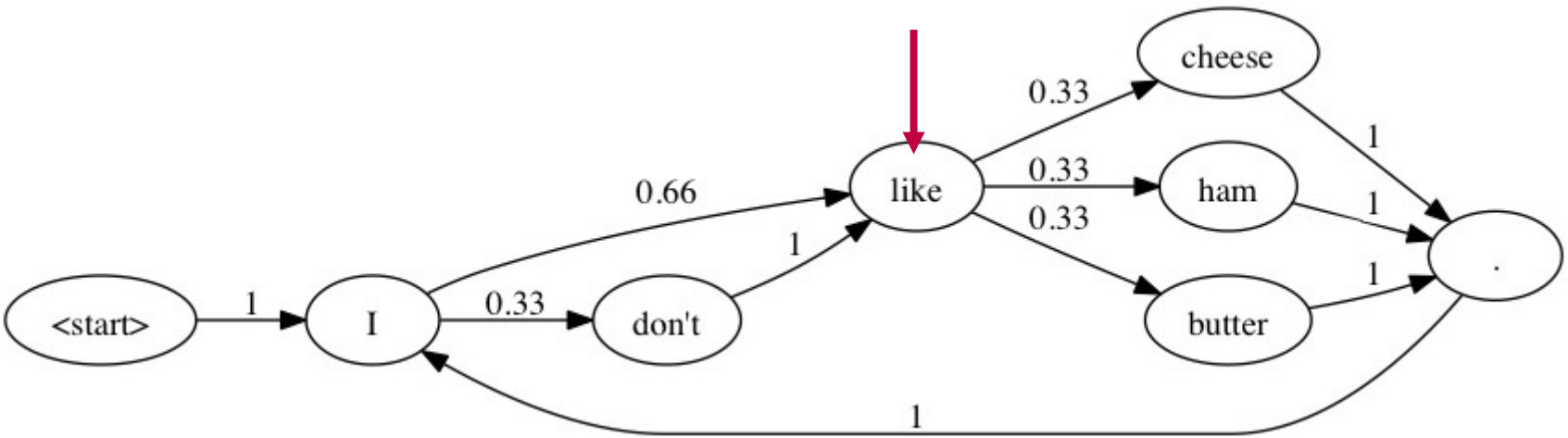


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “I don’t like”

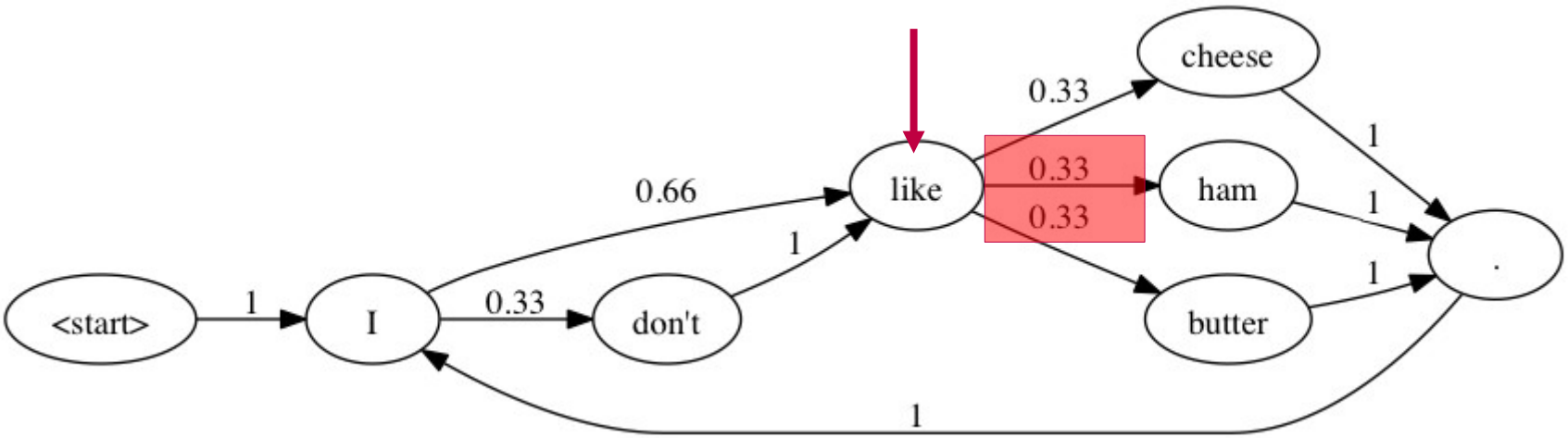


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “I don’t like ham”

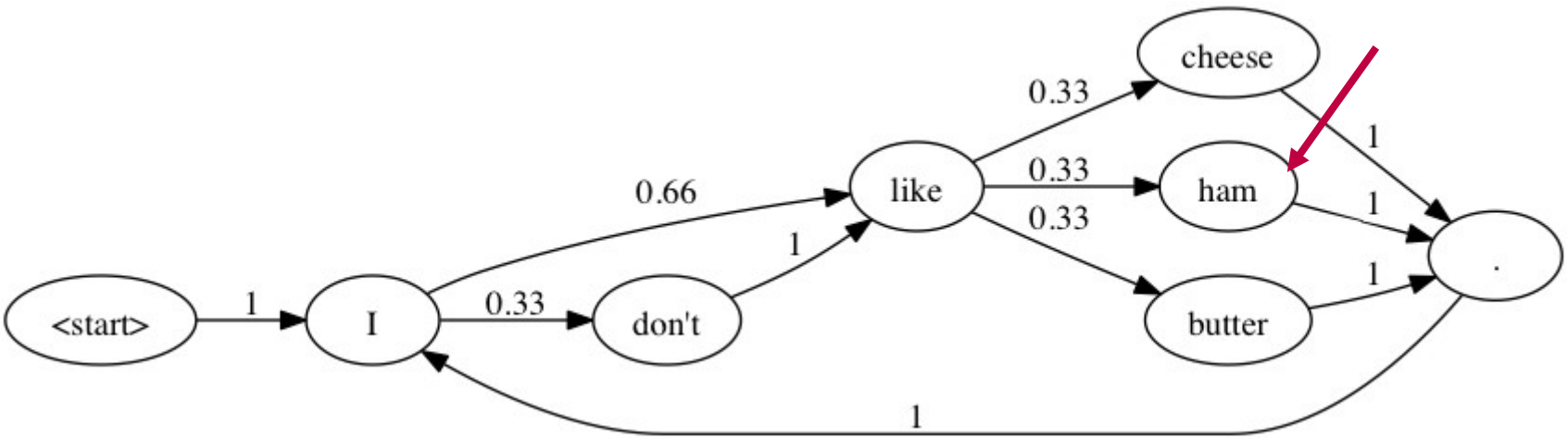


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “I don’t like ham”

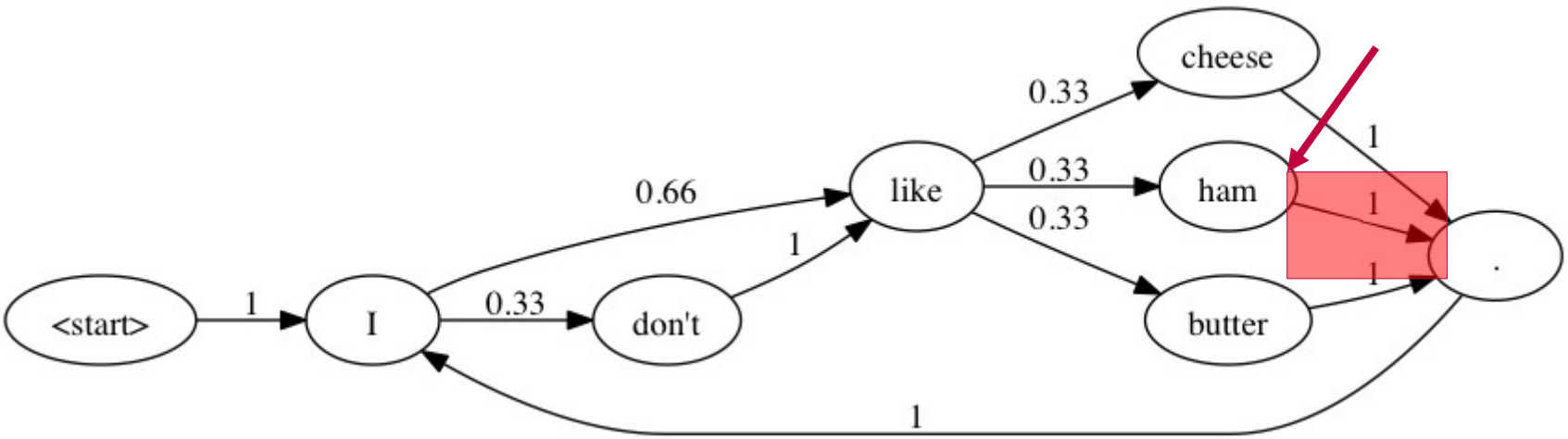


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

Our sentence: “I don’t like ham.”

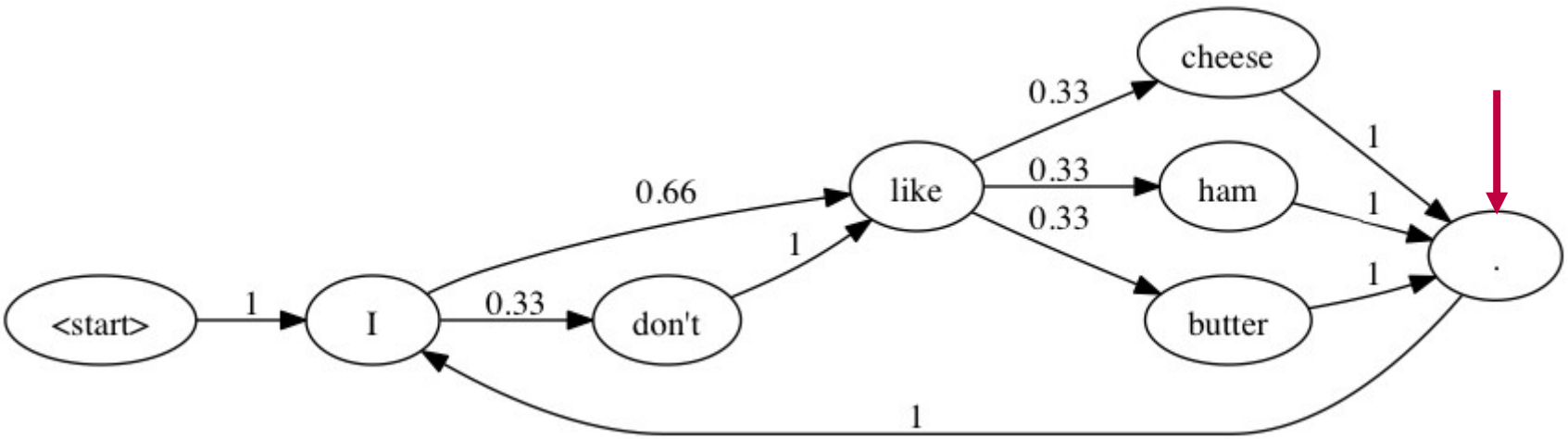


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

- Pedantic note: this graph has no <end> state!
  - But imagine it had one

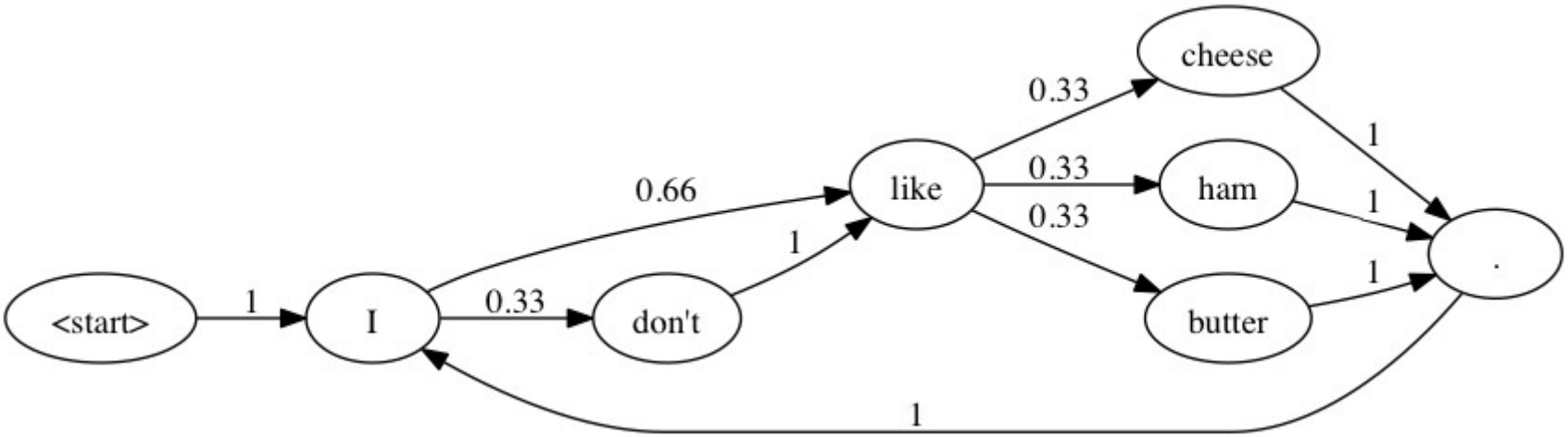


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

- We can build a graph like this from a dataset by computing the *transition* probabilities over a *corpus*

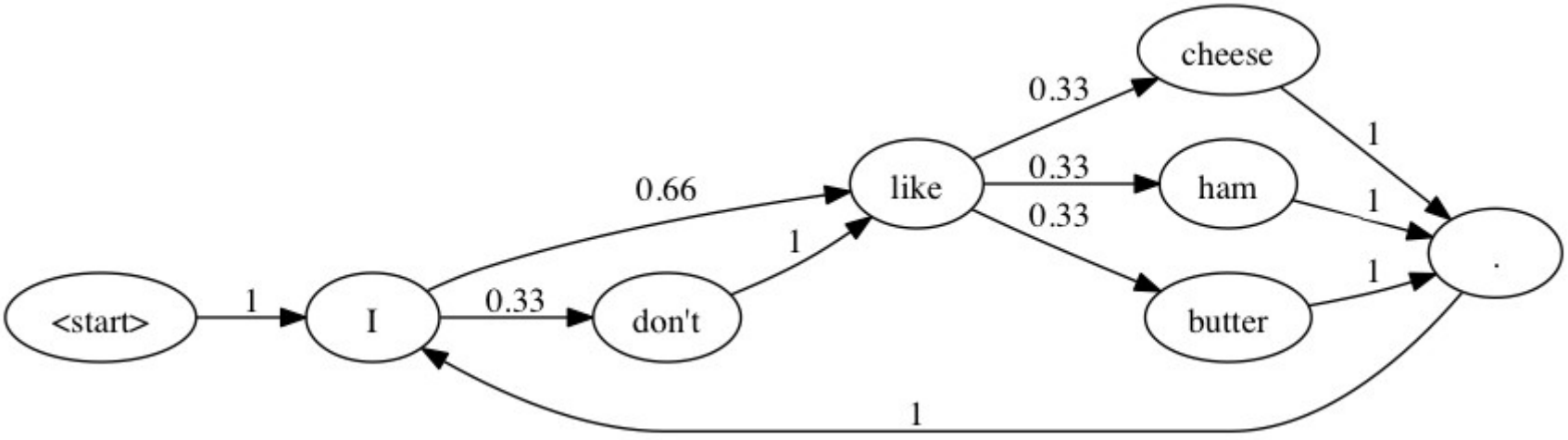


Figure from “The Making of Project Haikuza: Part 2”
https://www.justinmklam.com/posts/2015/making-haikuza-ii/

N-gram models

MY N-gram model

MY **N-gram model**

- https://github.com/SnoopJeDi/hn_markov
- Built with https://github.com/jsvine/markovify
- Corpus built by scraping some posts and comments from HackerNews