Capabilities of Markov models for tweet generation

Jamie Hackney

Goal

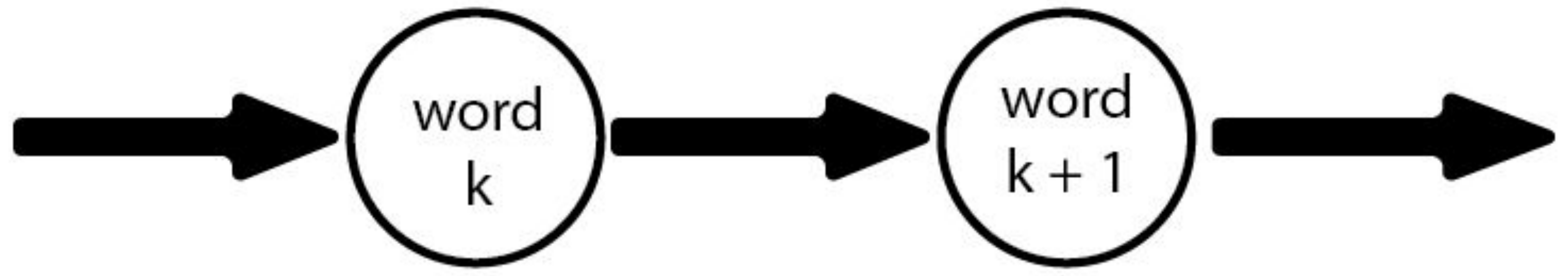
The goal of this project was to explore the capabilities of Markov models to generate tweets. Specifically, this project looked at 3 different Markov models and their generative capabilities.

Background

The three models are all variants of Markov chains:

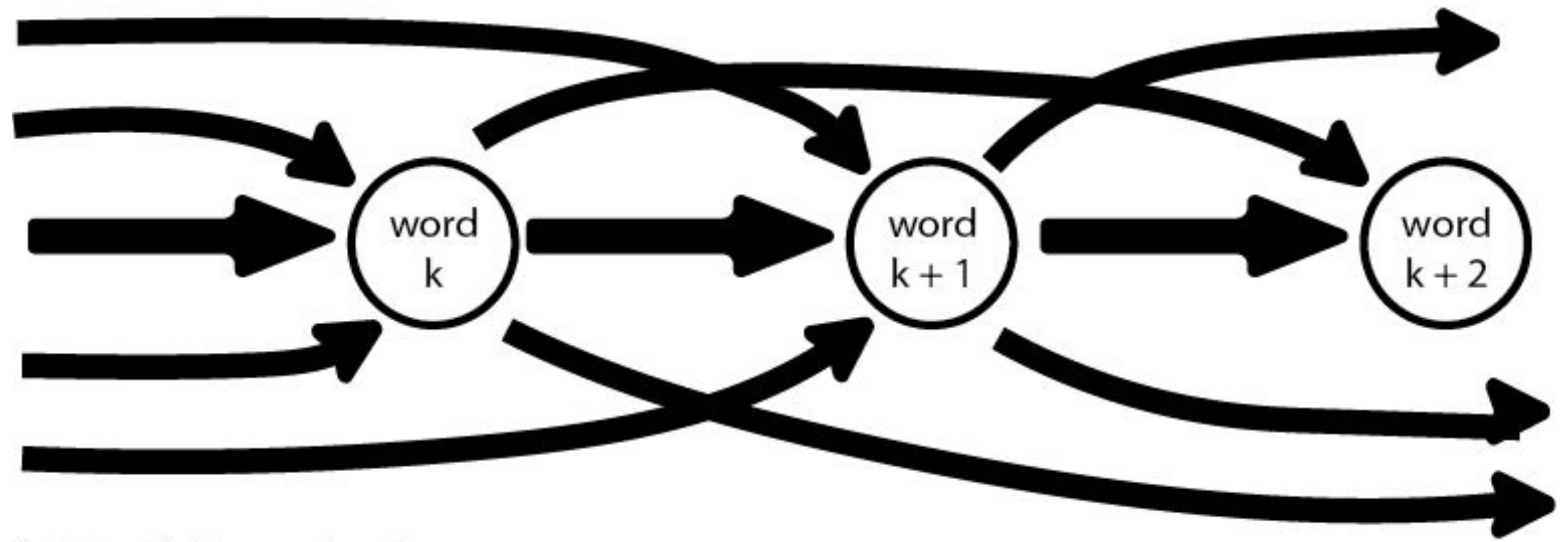
V1 Model

First Order Markov Chain



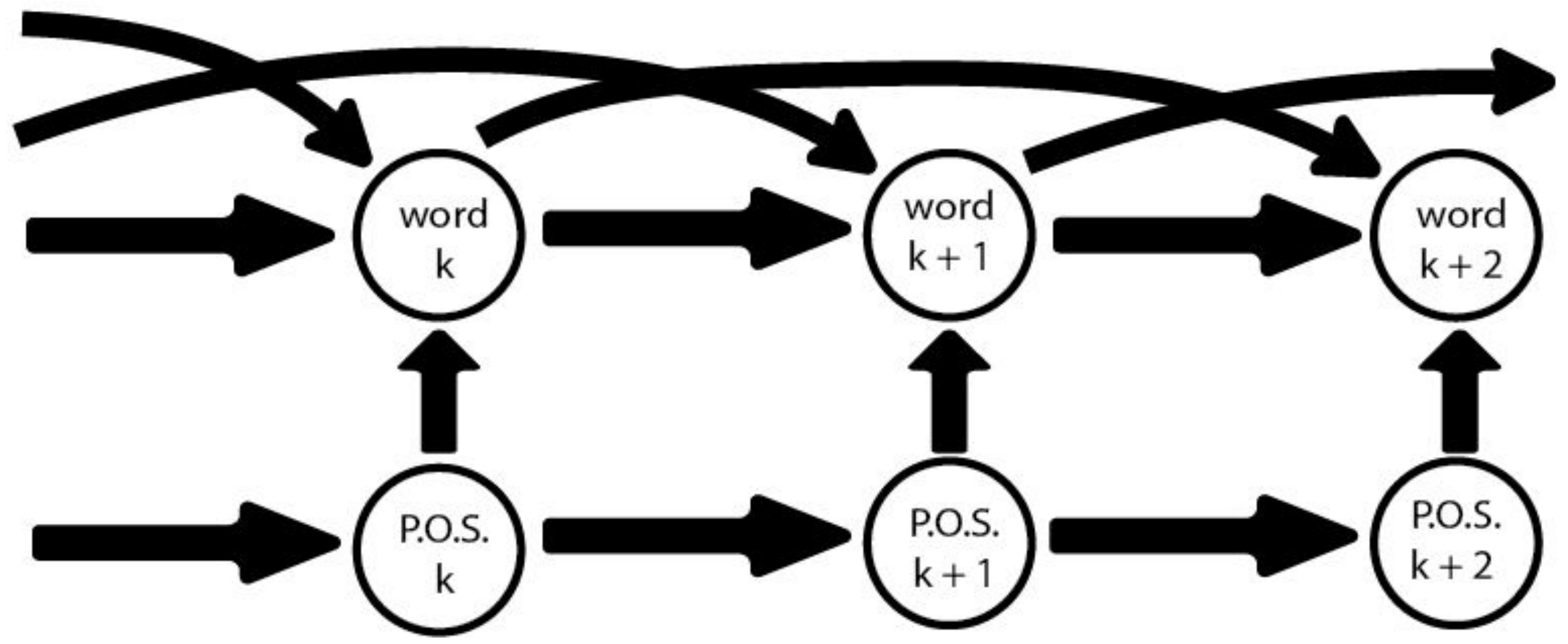
V2 Model

Third Order Markov Chain



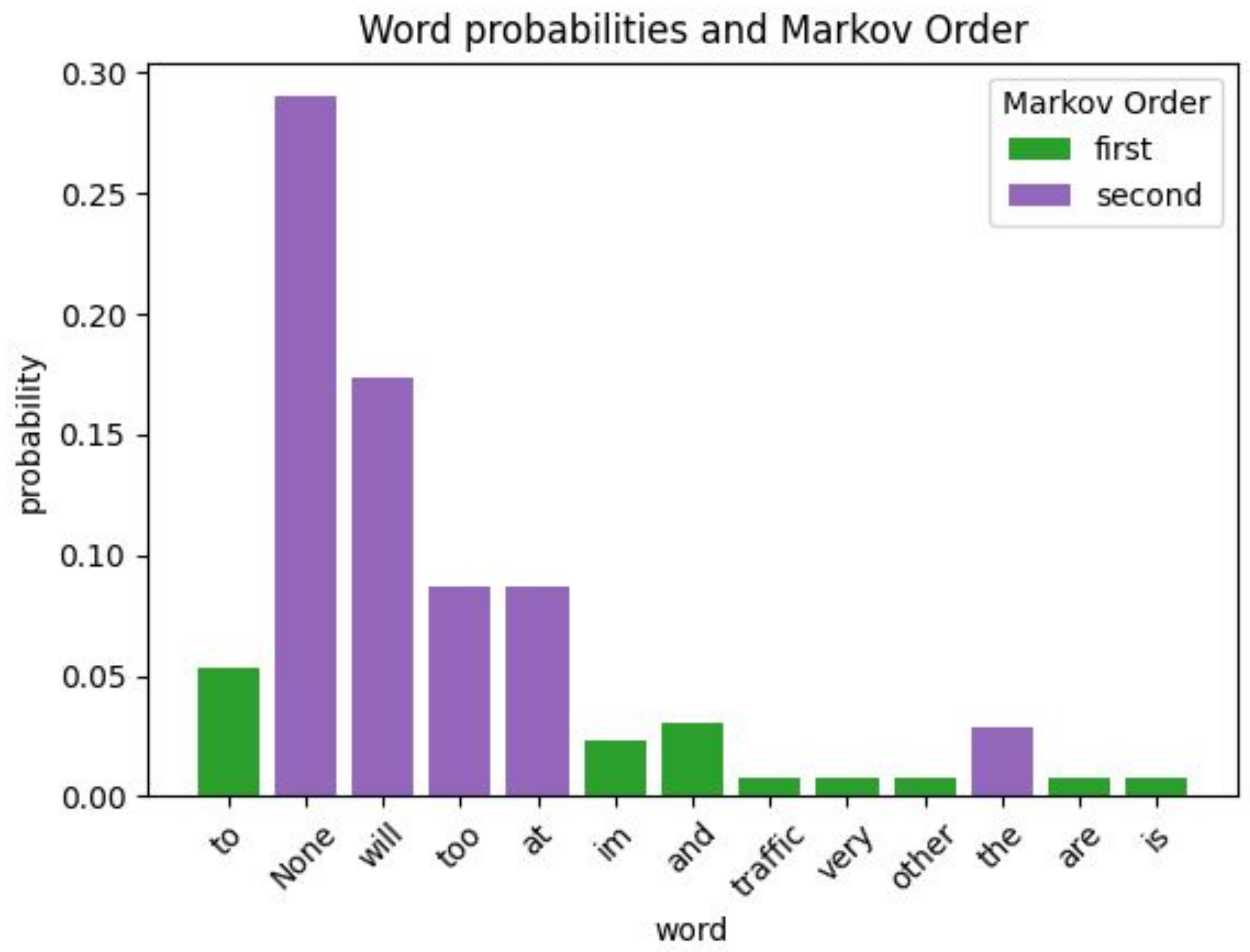
V3 Model

Second Order Markov Chain with P.O.S. tagging



Methods

These models work by calculating the probability distribution for the next word, and then sampling from that distribution. In the case of the V2 and V3 models, this distribution is a weighted combination of the first, second, and third order probabilities.



Above is the full probability distribution for the string "customers in canada" for the V2 model trained on Elon Musk tweets.

- Green bars represent first order probabilities
  - Pr(w | "*canada*")
- Purple bars represent second order probabilities
  - Pr(w | "*in canada*")
- Red (not present) would represent third order probabilities
  - Pr(w | "*customers in canada*")
  - Since the string "customers in canada" was not present in any training tweets, the model has no distribution for it.

Results

The models were trained on 5 data sets: Trump Tweets, Biden Tweets, Elon Musk Tweets, and Democrat and Republican tweets about the 2020 election. In general, the shorter generated tweets tend to be more coherent.

V1:

This model performed the worst of the 3 in its ability to synthesize coherent and grammatically correct tweets:

(democrats) we...

we continue in open positions my fifth most hardworking american journalists who was done to meet up for scoring 2

V2:

This model performed the best out of the 3, showing a capability to generate semi-coherent tweets:

(musk) we are going to mars...

we are going to mars to become multiplanetary

(biden) today i will...

today i will head to unleash the deepest darkest forces in this country that vows vengeance toward one friend @stephenathome

V3:

This model failed to consistently generate coherent tweets, likely due to a combination of:

- Lack of data limiting the possible word emissions
- Difficulties correctly tagging POS in often grammatically challenged tweets

(trump) we are...

we are a hoax

(trump) tonight we...

tonight we need @realdonaldtrump no surprise woman i #maga i love @realdonaldtrump trump agenda i will have him #trump2016