Lecture 4 Part 2:

n-grams

General probability of a word sequence

- According to the rules of joint probabilities and conditional probabilities, the probability of a observing a particular sequence of N words $w_1, w_2, w_3, \ldots, w_N$ can always be broken down like this:

$$p(w_1, w_2, w_3, \ldots, w_N) =$$
$$p(w_1)\, p(w_2|w_1)\, p(w_3|w_1, w_2)\, p(w_4|w_1, w_2, w_3) \cdots p(w_N|w_1, w_2, w_3, \ldots, w_{N-1})$$

- Each word can depend on all previous words in an arbitrary way
- While this probability distribution is completely general it is in practice impossible to learn
- Most statistical models used in text analytics make strong simplifying assumptions, we will discuss several
- **N-grams** assume words do not depend on all previous words, but only the nearby ones

LECTURE 4 PART 2:

n-GRAMS

Chapter 10: Unigrams

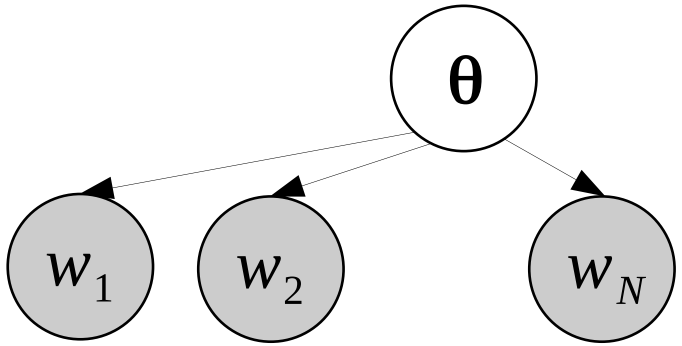
Unigram model

- The simplest n-gram model is the **unigram model**, also known as a **bag-of-words model**.
- Words are sampled fully independently of each other
- Each word is distributed according to a **multinomial distribution:** discrete distribution with W options, where W is the number of words in the vocabulary
- The parameter that defines the multinomial distribution is θ, a vector of word probabilities for every unique word in the vocabulary (the probabilities sum to 1):

$$\boldsymbol{\theta} = [\theta_{1,} \theta_{2,} \ldots, \theta_V]$$

Unigram model

- How to generate a document from the unigram model:
  - First select the number of words N (length of the document).
    - The unigram model does not define how the lengths of documents are generated, a separate model can be used for that.
  - For each word 1 to N, randomly sample a word from the multinomial distribution θ.



Unigram model

- **Probability of observations:**
- Consider a sequence of N words $w_{1,} w_{2,} w_{3,} \ldots, w_N$ where each w is an index into the vocabulary, e.g. $w_1 = 1$ means 'apple', $w_1 = 2$ means 'orange' and so on.
- Probability of the sequence in an unigram model:

$$p([w_{1,} w_{2,} w_{3,} \ldots, w_N] \ |\boldsymbol{\theta}) = p(w_1|\boldsymbol{\theta})\, p(w_2|\boldsymbol{\theta})\, p(w_3|\boldsymbol{\theta}) \cdots p(w_N|\boldsymbol{\theta})$$

$$= \prod_{i=1}^{N} p(w_i|\boldsymbol{\theta})$$

$$= \prod_{i=1}^{N} \theta_{w_i}$$

- The order of appearance of the words does not affect the probability of the sequence

Unigram model

- Probability that, in a sequence of N words, each unique word in the vocabulary $w_1, w_2, w_3, \ldots, w_V$ appears a particular number of times $n_1, n_2, n_3, \ldots, n_V$ (where the numbers sum to N):

$$p([n_1, n_2, \ldots, n_V] \; | \boldsymbol{\theta}) = \frac{N!}{n_1! n_2! \cdots n_V!} p(w_1|\boldsymbol{\theta})^{n_1} p(w_2|\boldsymbol{\theta})^{n_2} \cdots p(w_V|\boldsymbol{\theta})^{n_V}$$

$$= \frac{N!}{n_1! n_2! \cdots n_V!} \prod_{i=1}^{V} \theta_i^{n_i}$$

- The probability above is a sum over all possible orders of appearance of the words in a sequence of N words: the first term $\frac{N!}{n_1! n_2! \cdots n_V!}$ is the number of possible orders.

Unigram model

- **Estimation of the model:** Given a sequence of words $\boldsymbol{w}=[w_{1,} w_{2,} w_{3,} \ldots, w_N]$ assume they were generated by a unigram model and estimate its parameter $\boldsymbol{\theta}$
- **Maximum likelihood estimation:** find the parameter value that maximizes the probability of the word sequence.
- Count how many times (zero or more) each word in the vocabulary appears in the sequence: $n_1, n_2, n_3, \ldots, n_V$
- Result: $\boldsymbol{\theta}_{ML}=max_{\boldsymbol{\theta}} \; p([w_1, \ldots, w_N] \; |\boldsymbol{\theta})=\left[\frac{n_1}{N}, \ldots, \frac{n_V}{N}\right]$
- Once again, order of words did not matter
- **Several sequences**: same result, count words from all sequences

Unigram model

- **Maximum a posteriori (MAP) estimation:**
  - set a prior distribution for what probability vectors you believe are probable.
  - Then find the parameter value that has the highest posterior probability, according to the Bayes rule
- From the Bayes rule:

$$p(\boldsymbol{\theta}|\boldsymbol{w}) = \frac{p(\boldsymbol{w}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\boldsymbol{w})} = \frac{p(\boldsymbol{w}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}'} p(\boldsymbol{w}|\boldsymbol{\theta}')\,p(\boldsymbol{\theta}')}$$

$$\boldsymbol{\theta}_{MAP} = max_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta}|\boldsymbol{w}) = max_{\boldsymbol{\theta}}\, p(\boldsymbol{w}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})$$

- Example: choose a **Dirichlet prior** with pseudocounts $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_V]$

$$p(\boldsymbol{\theta}) = \frac{\Gamma\left(\sum_{i=1}^{V} \alpha_i\right)}{\prod_{i=1}^{V} \Gamma(\alpha_i)} \prod_{i=1}^{V} \theta_i^{\alpha_i}$$

- Result: $\boldsymbol{\theta}_{MAP} = \left[\frac{n_1 + \alpha_1}{N + \sum_i \alpha_i}, \ldots, \frac{n_V + \alpha_V}{N + \sum_i \alpha_i}\right]$

Unigram model

- **Full Bayesian posterior distribution:**
  - instead of one parameter value, infer a whole distribution of what parameters could have generated the data.
  - Set a prior distribution, get the posterior by the Bayes rule
  - If the prior is a Dirichlet distribution, it is a conjugate distribution to the multinomial observation probability, so the posterior can be computed analytically.

$$p(\boldsymbol{\theta}) = \frac{\Gamma\left(\sum_{i=1}^{V} \alpha_i\right)}{\prod_{i=1}^{V} \Gamma(\alpha_i)} \prod_{i=1}^{V} \theta_i^{\alpha_i}$$

- The posterior is another Dirichlet distribution:

$$p(\boldsymbol{\theta}|\mathbf{w}) = \frac{\Gamma\left(\sum_{i=1}^{V} \alpha_i^{posterior}\right)}{\prod_{i=1}^{V} \Gamma(\alpha_i^{posterior})} \prod_{i=1}^{V} \theta_i^{\alpha_i^{posterior}}$$

where the pseudocounts are

$$\boldsymbol{\alpha}^{posterior} = [\alpha_1^{posterior}, \ldots, \alpha_V^{posterior}] = [\alpha_1 + n_1, \ldots, \alpha_V + n_V]$$

mean:

$$E_{p(\boldsymbol{\theta}|\mathbf{w})}[\boldsymbol{\theta}] = \boldsymbol{\theta}_{MAP} = \left[\frac{n_1 + \alpha_1}{N + \sum_i \alpha_i}, \ldots, \frac{n_V + \alpha_V}{N + \sum_i \alpha_i}\right]$$

The N-gram Adventures of Sherlock Holmes

Part 1:

The Mystery of the Bag of Words

Unigram model

- Use Project Gutenberg text "The Adventures of Sherlock Holmes" (Project Gutenberg e-text) to learn a unigram model, with words lowercased but no other preprocessing (no lemmatization or pruning)
- Sample a string of words from a unigram model as described on the slide "How to generate a document from the unigram model"
- **Result 1:** red the wonder . there . of all path . from burned sort , . he deserts twisted . you out very fear am to of cylinders i put years and been with liability i myself unique be door , all in ! expensive goes would a . me the . a door miss hardly its like , them the into . you pen did which husband i shall opened almost of . she was . candle regency but sprang . were took hands passed thing conveyed however told . i you recommence am the of holder and sure but
- **Result 2:** the of concealed should but within giant . is our same . at the a dashed which . . at in , a a door fears of implore a and i the . sinister queen was of table to heart is briony beer his and colonies led my mr my of grotesque perform looking can we day said , , , disguised are , do engaged had story , is , i slow i the here from it smoking station ! and , wing sir a only one three 100 . perhaps impression came prevent baboon cargo complex a
- **Result 3:** art indignation strike think hunter of so doctor , , it away did i to hence may first dr is listened his , say but the clear guilty the once and , couple in frockcoat languid dazed . bounded are leave threatens contributions sound thoughtfully holmes a could lips him also ‘you holmes and the society to heart—it . it no that , seats proceed be in were , the of her front , the a lie he of from i trapdoor surely so size out midday , strange . describe charming comply and one , matter were the

Lecture 4 Part 2:

n-grams

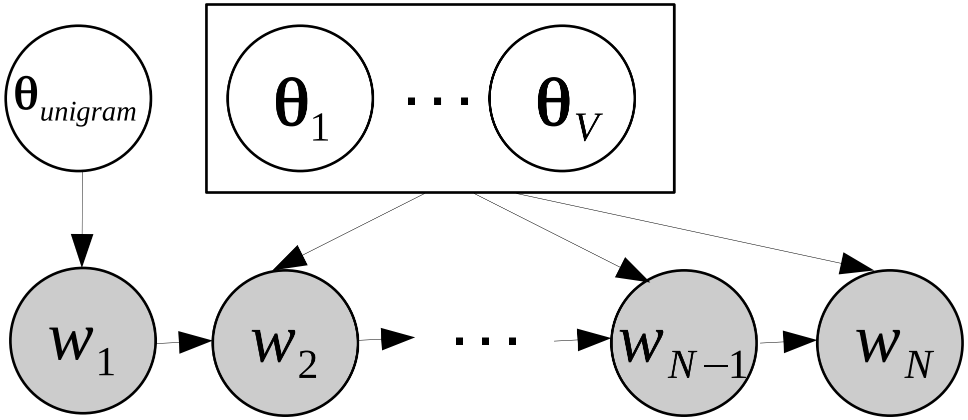
Chapter 11: Bigrams

Bigram model

- In a unigram model "dog bites man" and "man bites dog" are equally likely, because words are independent and their order does not matter
- The **bigram model** is the simplest n-gram model where words are not independent.
- Words are sampled depending on the previous word
- Each word is distributed according to a multinomial distribution, but its parameter depends on the previous word.
- Each word w in the vocabulary has its own multinomial parameter (probability vector) $\boldsymbol{\theta}_w$ that tells what words are likely to follow next: $\boldsymbol{\theta}_w = [\theta_{1|w}, \theta_{2|w}, \ldots, \theta_{V|w}]$

Bigram model

- How to generate a document from the bigram model:
  - Choose the number of words N
  - Generate the first word w_1 from a unigram model
  - Repeat for $i=2,\ldots,N$: if the previous word (i-1) has vocabulary index w_{i-1}, generate word n from the multinomial distribution $p\left(w|\boldsymbol{\theta}_{w_{i-1}}\right)$



Bigram model

- **Probability of observations:**
- Consider **several word sequences** $\boldsymbol{w}^{(s)}, \quad s=1,\ldots,S$ where each has $N^{(s)}$ words $\boldsymbol{w}^{(s)}=[w_1^{(s)}, w_2^{(s)}, w_3^{(s)}, \ldots, w_{N^{(s)}}^{(s)}]$ (they are again indices into the vocabulary)
- Probability of the sequences in a bigram model:

$$p(\boldsymbol{w}^{(1)},\ldots,\boldsymbol{w}^{(S)} \mid \boldsymbol{\theta}_{unigram},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_V)=\prod_{s=1}^{S} p(\boldsymbol{w}^{(s)} \mid \boldsymbol{\theta}_{unigram},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_V)$$

$$=\prod_{s=1}^{S} p(w_1^{(S)} \mid \boldsymbol{\theta}_{unigram})\, p(w_2^{(S)} \mid \boldsymbol{\theta}_{w_1})\, p(w_3^{(S)} \mid \boldsymbol{\theta}_{w_2}) \cdots p(w_{N^{(S)}}^{(S)} \mid \boldsymbol{\theta}_{w_{N^{(S)}-1}})$$

$$=\prod_{s=1}^{S} p(w_1^{(S)} \mid \boldsymbol{\theta}_{unigram}) \prod_{i=2}^{N^{(S)}} p(w_i \mid \boldsymbol{\theta}_{w_{i-1}^{(S)}})$$

$$=\prod_{s=1}^{S} \theta_{w_1^{(S)} \mid unigram} \prod_{i=2}^{N^{(S)}} \theta_{w_i^{(S)} \mid w_{i-1}^{(S)}}$$

- The word order now affects the probability of each sequence

Bigram model

- **Maximum likelihood estimation:**
  - count how many times, over all sequences, each vocabulary word appears in the sequence **before another word** (i.e. not at the end): $n_1, n_2, n_3, \ldots, n_V$
  - count how many times each word appears **in the beginning of a sequence**: $n_{1|unigram}, \ldots, n_{V|unigram}$
  - count how many times each word appears after each other word: $n_{1|1}, n_{2|1}, \ldots, n_{V|1}, \ldots, n_{1|V}, n_{2|V}, \ldots, n_{V|V}$
  - Result:

$$\boldsymbol{\theta}_{ML,unigram} = \left[\frac{n_{1|unigram}}{S}, \ldots, \frac{n_{V|unigram}}{S}\right] \quad , \quad \boldsymbol{\theta}_{ML,w} = \left[\frac{n_{1|w}}{n_w}, \ldots, \frac{n_{V|w}}{n_w}\right]$$

Bigram model

- **Maximum a posteriori estimation**: set independent Dirichlet priors
  - ... for $\boldsymbol{\theta}_{unigram}$ with pseudocounts $\boldsymbol{\alpha}_{unigram}=[\alpha_{1|unigram},\ldots,\alpha_{V|unigram}]$
  - ... for each $\boldsymbol{\theta}_w$ with pseudocounts $\boldsymbol{\alpha}_w=[\alpha_{1|w},\ldots,\alpha_{V|w}]$
  - Result:

$$\boldsymbol{\theta}_{MAP,unigram}=\left[\frac{n_{1|unigram}+\alpha_{1|unigram}}{S+\sum_i \alpha_{i|unigram}},\ldots,\frac{n_{V|unigram}+\alpha_{V|unigram}}{S+\sum_i \alpha_{i|unigram}}\right]$$

$$\boldsymbol{\theta}_{MAP,w}=\left[\frac{n_{1|w}+\alpha_{1|w}}{n_w+\sum_i \alpha_{i|w}},\ldots,\frac{n_{V|w}+\alpha_{V|w}}{n_w+\sum_i \alpha_{i|w}}\right]$$

Bigram model

- **Full Bayesian posterior**:
  - Set again independent Dirichlet priors.
  - Likelihood can be written as a product of independent terms for each parameter vector:

$$p\left(\boldsymbol{w}^{(1)},\ldots,\boldsymbol{w}^{(S)}\ |\boldsymbol{\theta}_{unigram},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_V\right)=\left(\prod_{i=1}^{V}\theta_{i|unigram}^{n_{i|unigram}}\right)\prod_{m=1}^{V}\left(\prod_{k=1}^{V}\theta_{k|m}^{n_{k|m}}\right)$$

  - Like the prior and likelihood, the posterior is a product of independent Dirichlet distributions for each parameter vector: $p\left(\boldsymbol{\theta}_{unigram}|\boldsymbol{w}^{(1)},\ldots,\boldsymbol{w}^{(S)}\right)\prod_{i=1}^{V}p\left(\boldsymbol{\theta}_i|\boldsymbol{w}^{(1)},\ldots,\boldsymbol{w}^{(S)}\right)$
  - Result:

$$p\left(\boldsymbol{\theta}|\boldsymbol{w}^{(1)},\ldots,\boldsymbol{w}^{(S)}\right)=\frac{\Gamma\left(\sum_{i=1}^{V}\alpha_i^{posterior}\right)}{\prod_{i=1}^{V}\Gamma\left(\alpha_i^{posterior}\right)}\prod_{i=1}^{V}\theta_i^{\alpha_i^{posterior}}$$

where for $\boldsymbol{\theta}_{unigram}$

$$\boldsymbol{\alpha}_{posterior,unigram}=\left[n_{1|unigram}+\alpha_{1|unigram},\ldots,n_{V|unigram}+\alpha_{V|unigram}\right]$$

and for each $\boldsymbol{\theta}_w$

$$\boldsymbol{\alpha}_{posterior,w}=\left[n_{1|w}+\alpha_{1|w},\ldots,n_{V|w}+\alpha_{V|w}\right]$$


The N-gram Adventures
of Sherlock Holmes
Part 2:
One Word
Leads to Another
ELSIE
MEET THY GOD

Bigram model

- Use "The Adventures of Sherlock Holmes" again to learn a bigram model, with words lowercased but no other preprocessing (no lemmatization or pruning)
- Sample a string of words from the bigram model as described on the slide "How to generate a document from the bigram model"
- Result 1: drifted into the palm of damages . i believe that she has nerve and grinning at the other people on my night . 8 or conscience . it was no trace . i will avail , taking the strange creature weaker than an ordinary plumber's smokerocket from any one owns a few years and put on the matter but i never had he paced about the business there is just transferred the strange disappearance . when he is said nothing definite result . she had solved in another woman . yes , fastening upon the incident of my pence every
- Result 2: him . but i was informed that i left . goodbye , you whether the morning , that all . watson , as you to their escape every prospect that you see , and gentle . it is my excuses , was struck , but i thought of white letters from his room early enough what i have seen anything which led the table . holmes nodded and passed his fingertips together . all these parts . of nitrate of the middle height , for me . lestrade would only all the sole in my afghan campaign throbbed with the
- Result 3: upon his eyes travelled in the redheaded man in particular shade of his manner suggested at last night . oh , i saw nothing actionable , and rushed down with an expenditure as being found it does not agree that i took to tell her father struck a cat , our whims so bound by that that the tail . terrible misfortune should feel better not trouble ! great public , upon the quick and left his features . the table , he heard a high , of his doings : some slight , and hurried from what day .