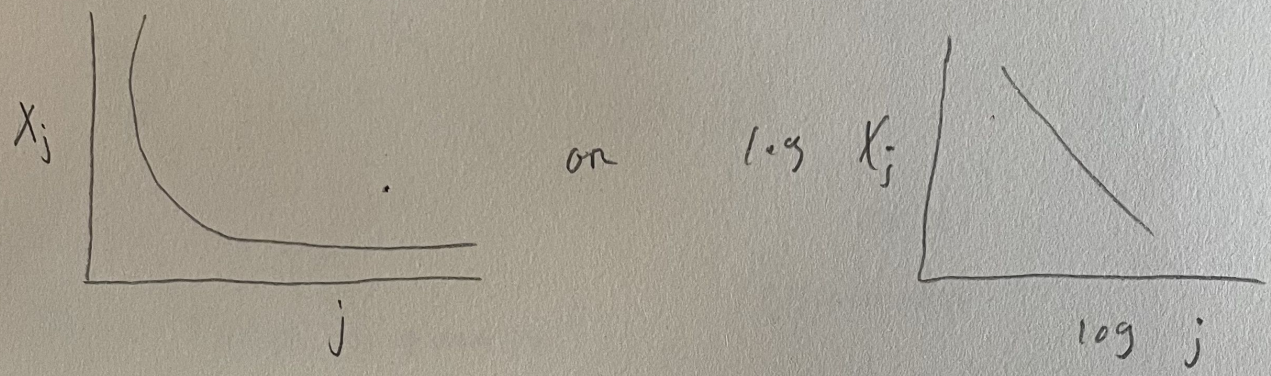
DISTRIBUTIONAL SEMANTICS P1

- MANY PROBLEMS IN NLP REQUIRE US TO ASSIGN A MEASURE OF SIMILARITY, OR RELEVANCE, BETWEEN PIECES OF TEXT:
  - WORDS
  - DOCUMENTS
  - CORPORA
- KEY PIECES OF TECHNOLOGY REQUIRE THIS:
  - INFORMATION RETRIEVAL / SEARCH SYSTEMS
  - RECOMMENDER SYSTEMS
  - VIRTUAL ASSISTANTS / CHATBOTS
- THUS FAR IN THIS COURSE WE HAVE USED WORD FREQUENCY (A.K.A. BOW A.K.A. WORD HISTOGRAMS) TO REPRESENT PIECES OF TEXT. THIS IS PROBLEMATIC:
  - OUR CORPS $X \in \mathbb{Z}^{M \times N}$ IS SPARSE
  - WE DON'T HAVE A RELIABLE DISTANCE MEASURE IN OUR $\mathbb{Z}^N$ INPUT VECTOR SPACE

- ANOTHER PROBLEM IS THAT P(X) IS HIGHLY NON-UNIFORM

→ Real ZIPF'S LAW: $P(X_j) = \frac{1/X_j^{\alpha}}{\sum_{j'=1}^{N} 1/j'^{\alpha}}$ $\alpha > 1$



→ IN MODULE #1 we LEARNED THAT "SELF INFORMATION" IS PROPORTIONAL TO THE INVERSE LOG OF P(X): $I(X) = -\log P(X)$

- HIGH FREQUENCY WORDS CARRY LOW INFORMATION
- LOW FREQUENCY WORDS CARRY HIGH INFORMATION

→ MANY EARLY IR SYSTEMS WERE DESIGNED TO BE SENSITIVE TO LOW FREQUENCY WORDS, AND INSENSITIVE TO HIGH FREQUENCY WORDS.

→ THIS CAN BE ACCOMPLISHED BY WEIGHING THE WORD COUNTS IN X ACCORDING TO FREQUENCY

TF-IDF ALGORITHM

- TF-IDF MAPS FREQUENCY COUNTS IN X ONTO AN ASSOCIATED SET OF "WEIGHTS" ACCORDING TO:

$$W_{ij} = TF(X)_{ij} \cdot IDF(X)_j$$

- TF := "TERM-FREQUENCY"

$$TF(X)_{ij} = X_j^{(i)}$$

- IDF := "INVERSE DOCUMENT FREQUENCY"

$$IDF(X)_j = \left[\sum_{i=1}^{M} \mathbb{1}\{ X_j^{(i)} > 0 \} \right]^{-1}$$

- MANY VARIANTS OF THE ABOVE. THE MOST WIDELY USED IS:

$$W_{ij} = \left(1 + \log X_j^{(i)}\right) \cdot \log\left(\frac{M}{\sum_{i'=1}^{M} \mathbb{1}\{ X_j^{(i')} > 0 \}} \right)$$

- DOCUMENT SEARCH USING TF-IDF

- GIVEN A TF-IDF WEIGHTED COUNTS MATRIX, $W \in \mathbb{R}^{M \times N}$, AND A QUERY, $q \in \mathbb{Z}^N$, OF WORD FREQUENCY COUNTS, FIND THE MOST RELEVANT DOCUMENT IN X (i.e., THE MOST RELEVANT ROW OF X)

- LET THE SIMILARITY BETWEEN DOCUMENT i AND OUR QUERY BE $S_i(q)$

$$S_i(q) = \sum_{j=1}^{N} 1\{q_j > 0\} \cdot W_{ij}$$

- THE MOST RELEVANT DOCUMENT IS THEN SELECTED ACCORDING TO:

$$\hat{i} = \underset{i \in \{1, \ldots, M\}}{\text{ARGMAX}} \; S_i(q)$$

POINTWISE MUTUAL INFORMATION (PMI)

- CAN BE USED IN THE SAME WAY THAT WE USE TF-IDF FOR IR

- $PMI(i, j) = \log \frac{P(i,j)}{P(i)P(j)}$ (doc: i, word: j)

$$= \log \frac{P(j|i)P(i)}{P(i)P(j)}$$

$$= \log \frac{P(j|i)}{P(j)} = \log P(j|i) - \log P(j)$$

$$= \log \frac{X_j^{(i)}}{\sum_{j'=1}^{N} X_{j'}^{(i)}} - \log \frac{\sum_{i'=1}^{M} X_j^{(i')}}{\sum_{i'=1}^{M}\sum_{j'=1}^{N} X_{j'}^{(i')}}$$

$$= \log\left(\frac{\text{COUNT}(i,j)}{\text{COUNT}(i, \text{ALL } j)}\right) - \log\left(\frac{\text{COUNT}(\text{ALL } i, j)}{\text{COUNT}(\text{ALL } i, \text{ALL } j)}\right)$$

- POSITIVE PMI (PPMI)

$$PPMI(i,j) = \begin{cases} PMI(i,j), & P(j|i) > P(j) \\ 0, & \text{ELSE} \end{cases}$$

- TF-IDF / PMI ARE USED TO ADDRESS PROBLEMS RELATED TO ZIPF'S LAW (i.e., THE NON-UNIFORMITY IN $P(x)$)
- THESE METHODS DO NOT CHANGE THE SPARSITY OF OUR REPRESENTATION, THOUGH!

- OUR GOAL IS TO MAP W ONTO A "DENSE" REPRESENTATION: $W \in \mathbb{R}^N \rightarrow U \in \mathbb{R}^K$, $K \ll N$

~~$X \in X' : \mathbb{R}^N \rightarrow \mathbb{R}^K$ where $K \ll N$~~

- THERE ARE TWO MATRIX FACTORIZATION METHODS THAT DO A GOOD JOB AT THIS IN PRACTICE
  - TRUNCATED SVD (
  - NON-NEGATIVE MATRIX FACTORIZATION

Latent Semantic Analysis (LSA)

- A.K.A. Latent Semantic Indexing (LSI)
- Uses Truncated SVD

- We start with a matrix of weighted Frequency Counts, $W \in \mathbb{R}^{M \times N}$

- Recall the Truncated SVD:

$$W = U \quad \Sigma \quad V^T$$

$M \times N \qquad M \times K \quad K \times K \quad K \times N$

↳ K is often chosen using
  - Elbow method performed on Σ from Full SVD
  - Just $K \ll N$

- Numerical packages in Python (sklearn, Scipy) can do this for you!

- Our compressed corpus is given by:

$$U = W V \Sigma^{-1}$$ → Recall that

V is orthogonal

Σ is a diagonal matrix

Document Search using LSA

- Using the Problem Statement From L4-5:

  - First Compute the weighted query vector: TF-IDF / PMI for example

    $\rightarrow q \Rightarrow q^{(w)}$

  - Project $q^{(w)} \in \mathbb{R}^N$ to $q'^{(w)} \in \mathbb{R}^K$

    $$q'^{(w)} = q^{(w)} V \Sigma^{-1}$$

  - Use Cosine-Similarity to measure the relevance of document $X^{(i)}$ to q

    $$S_i(q) = \cos(W^{(i)}, q'^{(w)})$$

    $$= \frac{\sum_{j=1}^{K} q'^{(w)}_j W^{(i)}_j}{\sqrt{\sum_{j=1}^{K} q'^{(w)2}_j} \sqrt{\sum_{j=1}^{K} W^{(i)2}_j}}$$

  - Document selection:

    $$\hat{i} = \underset{i \in \{1,\dots,M\}}{\operatorname{argmax}} S_i(q'^{(w)})$$

NON-NEGATIVE MATRIX FACTORIZATION (NMF)

- NOTE THAT IN LSA, THE TRANSFORMED VECTOR SPACE IN $\mathbb{R}^K$, IS DEFINED BY THE K-COLUMNS of V. IN OTHER WORDS, EACH COLUMN OF V REPRESENTS A BASIS VECTOR IN OUR NEW $\mathbb{R}^K$ SPACE. IMPORTANTLY, EACH of THESE BASIS VECTORS IS A LINEAR COMBINATION OF THE N-BASIS VECTORS IN THE ORIGINAL $\mathbb{R}^N$ SPACE. WITH LSA, THE COEFFICIENTS THAT GET APPLIED TO EACH ORIGINAL DIMENSION (I.E. WORD) CAN BE NEGATIVE, AND THIS IS UNDESIRABLE

  EXAMPLE: $V_j^T = (0.5)\text{SWIM} + (0.7)\text{SUN} + (-1.0)\text{RAIN} + (-0.2)\text{DARK}$

  COULD EQUIVALENTLY BE REPRESENTED AS:

  $(0.5)\text{SWIM} + (11.4)\text{SUN} + (0.01)\text{RAIN} + (0.01)\text{DARK}$

- NMF FACTORS X USING A POSITIVITY CONSTRAINT ON THESE COEFFICIENTS THAT GET APPLIED TO THE WORDS.

NMF Cont...

- NMF Factorization typically is performed on a weighted version of X, similar to LSA

$$W = UV \quad \Rightarrow \quad V \text{ is orthogonal}$$

$M \times N \quad M \times K \quad K \times N$

- Objective Function:

$$\hat{U}, \hat{V} = \underset{U,V \geq 0}{\min} \; L(W, U, V) + \alpha\rho(\|U\|_1 + \|V\|_1) + \alpha(1-\rho)\left(\|U\|_F^2 + \|V\|_F^2\right)$$

α = Regularization Constant

ρ = Sparsity Constant

L = Some objective function

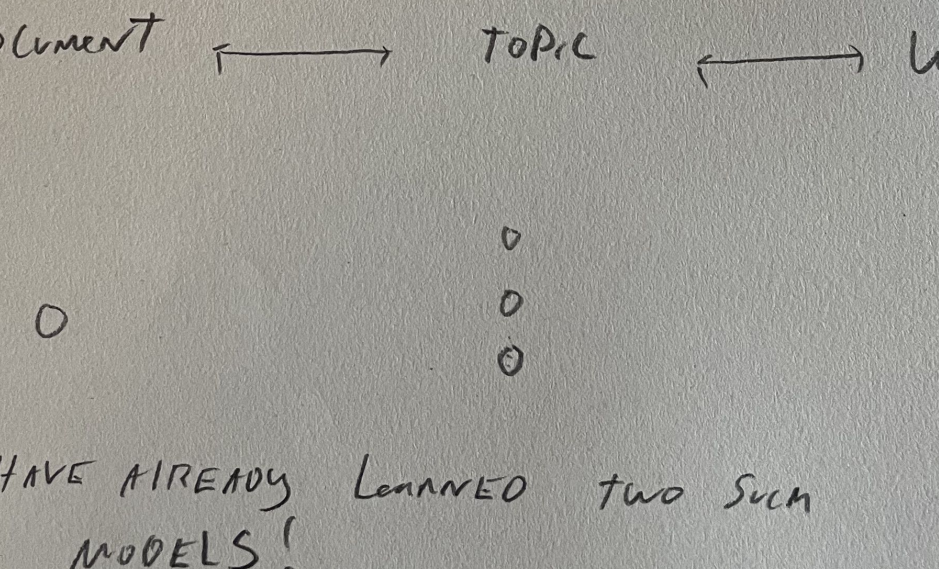
Frobenius Loss: $L_F = \|W - UV\|_F^2$

KL-Divergence Loss: $L_{KL} = \sum_{i=1}^{M} \sum_{j=1}^{N} W_{ij}^{(i)} \left[\log \frac{W_{ij}^{(i)}}{(UV)_{ij}} - 1 \right] + (UV)_{ij}$

TOPIC MODELING

- THUS FAR WE HAVE PRIMARILY BEEN CONCERNED WITH MODELING THE RELATIONSHIP BETWEEN DOCUMENTS AND THEIR CONSTITUENT WORDS.

- TOPIC MODELING IS A SET of METHODS THAT INVOLVE AN "INTERMEDIATE" REPRESENTATION THAT SEPARATES DOCUMENTS FROM WORDS. THE INTERMEDIARY IS REFERRED TO AS "TOPICS"



- WE HAVE ALREADY LEARNED TWO SUCH TOPIC MODELS!
  - LSA: $X = U \Sigma V^T$ → "TOPICS" = Columns of V
  - NMF: $X = UV$ → "TOPICS" = rows of V