Announcements – Assignments

- Readings 05:
  - No more this semester – congrats!

- HW02
  - Mostly done grading
  - You all did very well!!!

- HW04
  - Due Thursday
Announcements

- P/D/F deadline is today

- Course Evaluations
  - Due 06/14
Final Project – Deliverables

- Project ideation – Friday May 28\textsuperscript{st}  
  • 5 points

- Project proposal – Sunday June 6\textsuperscript{th}  
  • 9 points

- Project presentations – Monday June 14\textsuperscript{th}  
  • 6 points

- Project submissions – Friday June 18\textsuperscript{th}  
  • 15 points

http://coms2710.barnard.edu/final_project
5 minute presentations by each group

Format:
- Research Question
- Motivation
  - Why should we care?
- Data Collected
  - Where did the data come from? How did you collect it?
  - What filtering was done?
  - Resulting corpus:
    - How many documents? Average size of documents? Vocabulary size?
- Results (preliminary)
  - figures

Goal:
- Publicizing your research is important
- You all to see what everyone else is working on
“Congratulations, offer letter from Google”  Spam

“Congratulations, you won the lottery”  Not Spam

Which mistake is worse?
Evaluation Metrics

Recall
- When we do not want false negatives

Precision
- When we do not want false positives
Summary of Logistic Regression

- Optimizes $P(Y \mid X)$ directly
- Define the **features**
- Learn a vector of **weights** for each label $y \in Y$
  - Gradient descent, update weights based on error
- Multiple feature vector by weight vector
- Output is $P(Y = y \mid X)$ after normalizing
- Choose the most probable $Y$
Summary of Logistic Regression

- Optimizes $P(Y \mid X)$ directly
- Define the features
- Learn a vector of weights for each label $y \in Y$
  - Gradient descent, update weights based on error
- Multiple feature vector by weight vector
- Output is $P(Y = y \mid X)$ after normalizing
- Choose the most probable $Y$
Scoring one document - dot product

\[
[f_1, f_2, f_3] \cdot [w_1, w_2, w_3] = (f_1 \times w_1) + (f_2 \times w_2) + (f_3 \times w_3) = \sum (f_i \times w_i)
\]
Score two documents

\[
\begin{bmatrix}
[f_{1,1}, f_{1,2}, f_{1,3}] \\
[f_{2,1}, f_{2,2}, f_{2,3}]
\end{bmatrix}
\cdot
\begin{bmatrix}
w_1, \\
w_2, \\
w_3
\end{bmatrix}
\]

\[
= [(f_{1,1} \times w_1) + (f_{1,2} \times w_2) + (f_{1,3} \times w_3), \\
(f_{2,1} \times w_1) + (f_{2,2} \times w_2) + (f_{2,3} \times w_3)]
\]
Matrix Multiplication

We can multiply two matrices A and B if ....
number of columns in A = number of rows in B

The size of the resulting matrix is ....
number of rows in A & the number of columns in B
Matrix Multiplication

\[
\begin{bmatrix}
1 & 7 \\
2 & 4
\end{bmatrix}
\cdot
\begin{bmatrix}
3 & 3 \\
5 & 2
\end{bmatrix}
\]

\[A \quad B\]
Matrix Multiplication

\[
\begin{pmatrix}
1 & 7 \\
2 & 4
\end{pmatrix}
\begin{pmatrix}
3 & 3 \\
5 & 2
\end{pmatrix}
\]
Matrix Multiplication

\[
\begin{align*}
\vec{a}_1 & \rightarrow \begin{bmatrix} 1 & 7 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} 3 & 3 \\ 5 & 2 \end{bmatrix} = \\
\vec{a}_2 & \rightarrow \begin{bmatrix} \vec{a}_1 \cdot \vec{b}_1 & \vec{a}_1 \cdot \vec{b}_2 \\ \vec{a}_2 \cdot \vec{b}_1 & \vec{a}_2 \cdot \vec{b}_2 \end{bmatrix}
\end{align*}
\]
Question: What are the dimension of $C$?

2 rows x 2 columns
Document–Term Matrix
Document-Term Matrix

DMT:
- Rows represent a document
- Columns represent a word
- Values represent some feature of word \( w_i \) in document \( d_j \)

\[
\begin{array}{ccccccc}
\hline
w_1 & w_2 & w_3 & w_4 & \ldots & \ldots & \ldots & \ldots & w_v \\
\hline
\end{array}
\]

\[
\begin{array}{c}
d_1 \\
\hline
d_1 \\
\hline
\ldots \\
\hline
d_n \\
\hline
\end{array}
\]

10
Properties of Document-Term Matrix

- **Sparse matrix**
  - Most values are 0

- **Very large**
  - Many, many, many columns

- **Noisy**
Goal

- Make values in each cell more meaningful
- Reduce the size of the matrix
  - Dimensionality reduction
- Remove noise
Matrix Factorization/Dimensionality Reduction
Abstract thought

https://www.youtube.com/watch?v=dROx9Djr7mk
\[ M_{n \times v} = U_{n \times k} S_{k \times k} V^T_{k \times v} \]
Singular Value Decomposition

\[ M_{n \times v} = U_{n \times k} S_{k \times k} V_{k \times v} \]
Latent Semantic Analysis

- Applying SVD to the DTM is called Latent Semantic Analysis

- The name of LDA is based on this
• Monday 06/07 - Matrix Factorization

• Tuesday 06/08 – Word Embeddings

• Wednesday 06/09 – Guest Lecture
  • Attendance required

• Thursday 06/10 – ngrams & phrases