SMAI-S25–04: Data as Matrix

C. V. Jawahar

IIIT Hyderabad

January 17, 2025

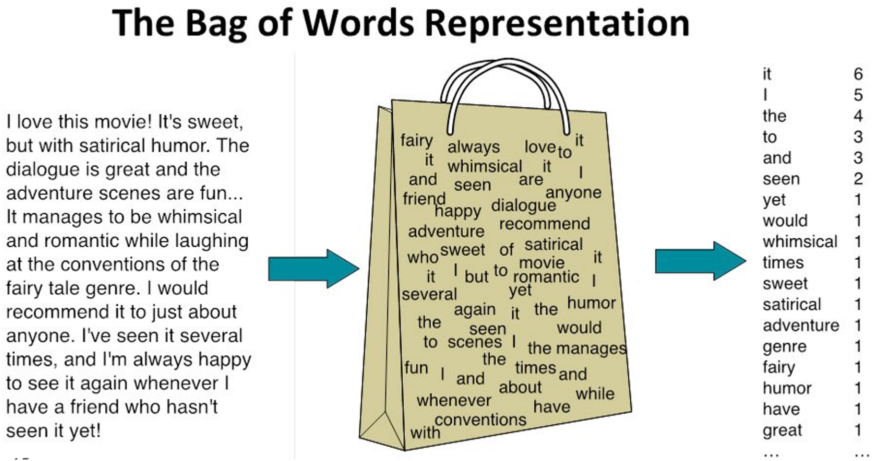
Administrative

1. Lectures
    - L01: https://www.dropbox.com/scl/fi/643pbhfworhj6nq2ks5gb/L01.pdf?rlkey=1k7e6tfvc0afd4z24bu684p9f&dl=0
    - L02: https://www.dropbox.com/scl/fi/n4xtyqqmdd26u03wiy0a7/L02.pdf?rlkey=59il9r4b0mgdydslv3v9xzvud&dl=0
    - L03:https://www.dropbox.com/scl/fi/yj0qv7pq00hu69dca9u80/L03.pdf?rlkey=g7d2k0ucub2pk00nmzx7c2qmy&dl=0
2. Logistics:
    - Project Teams:
        - https://docs.google.com/spreadsheets/d/1Qn5ot9ABVr0gG3u5dG4Qm2TMcgJb8qiV4rskX5fiZlY/edit?usp=sharing
        - Three in a team; form your team by Friday. (in rare cases 2 or 4! Look for 50+ teams) Also choose sports. And soon data.

Recap

1. Representation as a vector in R^d
   - All: Web page; Image; Song; Weather.
2. Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
   - Notion of Training and Testing (validation today)
3. Feature Transformation as a useful trick:
   - $\mathbf{x}' = f(\mathbf{W}, \mathbf{x})$; $\mathbf{x}' = \mathbf{W}\mathbf{x}$ and Dimensionality Reduction
4. Classification Algorithms:
   - Nearest Neighbour: Decide based on majority labels of K NNs
   - Linear Classification: Decide as ω_1 if $\mathbf{w}^T x \geq 0$ Else ω_2
   - **Next Lecture: Decide as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2**
5. Performance Metrics:
   - Classification: Accuracy, TP, FP etc., Confusion Matrix
   - Ranking: Precision, Recall, F-Score, AP

Bag of Words Histogram and One-Hot Representation



The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

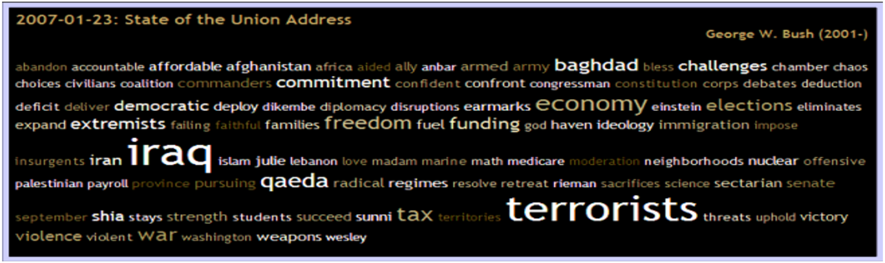
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

A simple exercise

Someone took all the speeches of past US presidents and created a Bag of Words Histograms. (one histogram per President). You can visualize the histogram as a word cloud (see next slide). Higher the frequency, the larger the font.

Q: Guess what it could be for Joe Baiden and Donald Trump. Who are the presidents in the next three slides? When were they (year?)

Representation: Bag of Words and One-Hot

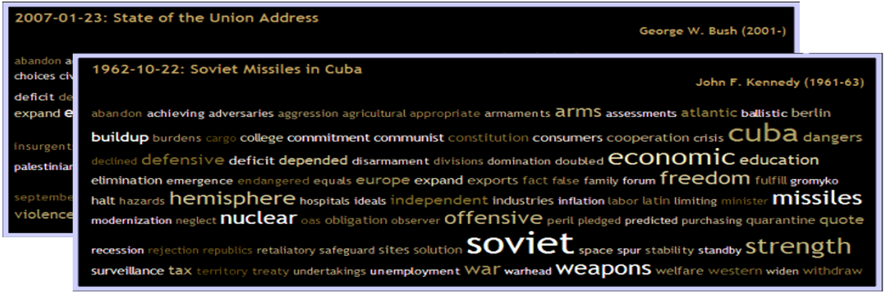


2007-01-23: State of the Union Address

George W. Bush (2001-)

abandon accountable affordable afghanistan africa aided ally anbar armed army baghdad bless challenges chamber chaos choices civilians coalition commanders commitment confident confront congressman constitution corps debates deduction deficit deliver democratic deploy dikembe diplomacy disruptions earmarks economy einstein elections eliminates expand extremists failing faithful families freedom fuel funding god haven ideology immigration impose insurgents iran iraq islam julie lebanon love madam marine math medicare moderation neighborhoods nuclear offensive palestinian payroll province pursuing qaeda radical regimes resolve retreat rieman sacrifices science sectarian senate september shia stays strength students succeed sunni tax territories terrorists threats uphold victory violence violent war washington weapons wesley

Representation: Bag of Words and One-Hot



2007-01-23: State of the Union Address

George W. Bush (2001-)

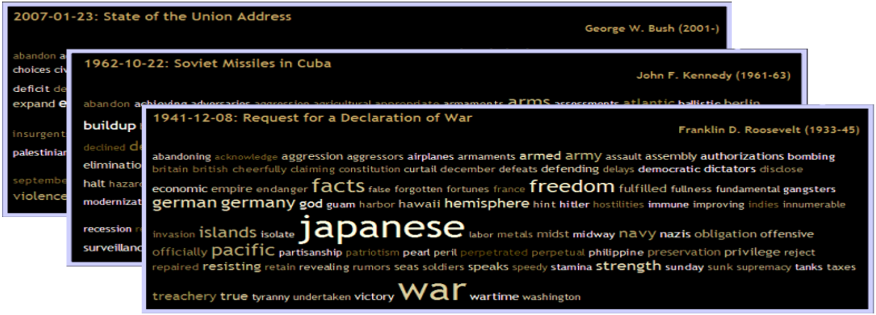
abandon a
choices civ
deficit de
expand e
insurgent
palestinian
septembe
violence

1962-10-22: Soviet Missiles in Cuba

John F. Kennedy (1961-63)

abandon achieving adversaries aggression agricultural appropriate armaments arms assessments atlantic ballistic berlin buildup burdens cargo college commitment communist constitution consumers cooperation crisis cuba dangers declined defensive deficit depended disarmament divisions domination doubled economic education elimination emergence endangered equals europe expand exports fact false family forum freedom fulfill gromyko halt hazards hemisphere hospitals ideals independent industries inflation labor latin limiting minister missiles modernization neglect nuclear oas obligation observer offensive peril pledged predicted purchasing quarantine quote recession rejection republics retaliatory safeguard sites solution soviet space spur stability standby strength surveillance tax territory treaty undertakings unemployment war warhead weapons welfare western widen withdraw

Representation: Bag of Words and One-Hot



2007-01-23: State of the Union Address

George W. Bush (2001-)

abandon a
choices civ
deficit de
expand e
insurgent
palestinian
septembe
violence

1962-10-22: Soviet Missiles in Cuba

John F. Kennedy (1961-63)

abandon achieving adversaries aggression agricultural appropriate armaments arms assessments atlantic ballistic berlin
buildup
declined de
eliminatio
halt hazard
modernizat
recession r
surveillanc

1941-12-08: Request for a Declaration of War

Franklin D. Roosevelt (1933-45)

abandoning acknowledge aggression aggressors airplanes armaments armed army assault assembly authorizations bombing
britain british cheerfully claiming constitution curtail december defeats defending delays democratic dictators disclose
economic empire endanger facts false forgotten fortunes france freedom fulfilled fullness fundamental gangsters
german germany god guam harbor hawaii hemisphere hint hitler hostilities immune improving indies innumerable
invasion islands isolate japanese labor metals midst midway navy nazis obligation offensive
officially pacific partisanship patriotism pearl peril perpetrated perpetual philippine preservation privilege reject
repaired resisting retain revealing rumors seas soldiers speaks speedy stamina strength sunday sunk supremacy tanks taxes
treachery true tyranny undertaken victory war wartime washington

Discussion Point

We consider 100 documents each from “sports” “politics” and “finance” and create a representation of size 300.

1. We construct a Data Matrix D by keeping each vector as a row. What is the dimension of this matrix? (simple!). Any need to normalize this histogram?
2. What could be the rank of this matrix? (wait!!. let us answer the next two questions first.)
3. Assume there was a small error in the code in that created this matrix. i.e., a single sports document was copied 100 times instead of different 100 articles (and similarly for polictics and finance). What will be the rank of the data matrix?
4. Assume a different situation. It was only one sports article; and it was ”rewritten” by 100 journalists (like some plagiarism!). What will be the rank of the data matrix?

Revise

Revise your understanding on the following topics:

- Matrices and Properties
- Eigen values and Eigen vectors
- Determinants
- Rank
- SVD and Matrix Decompositions (Advanced!)

before the next lecture.

Book: `https://mml-book.github.io/` Most of Chapter 2 and Chapter 4.

Discussion Point

Without numerically computing, what is the rank of the following matrix?

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

1. 1
2. 2
3. 3
4. 4

Hint:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 1+3 & 2+3 & 3+3 \\ 1+6 & 2+6 & 3+6 \end{bmatrix}$$

Discussion Point

Consider a matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} + \epsilon B$$

where B is a full rank matrix formed by random integers in 1 to 100.

- When $\epsilon = 1$, what is the rank of A?
- When $\epsilon = 10^{-100}$, what is the rank of A?
- How many "non-zero" eigen values will A have? Can you comment on their magnitude in both cases?

Problem I

A big sporting company has a data on good days to play cricket outdoor. An SMAI student (Raju) implements $K - NN$ and want to provide a solution. He conducted an experiment to vary K (say from 3 to 15) and plot the performance.
Q1:

- Will he see a systematic increase in accuracy with K?
- Will he see a systematic decrease in accuracy with K?
- Will he see a systematic increase followed by a systematic decrease?

Q2: Can you help Raju in finding the best K?

Problem-II

Q: Consider a linear transformation $d \rightarrow d$ (i.e., $\mathbf{W}$ is a square matrix)

$$\mathbf{x}' = \mathbf{W}\mathbf{x}$$

We use a K-NN algorithm (the same K and distance as Euclidean distance) in original and new space.

- Will the performance (say accuracy) of the algorithm be same in both the space for any $\mathbf{W}$? i.e., with $\mathbf{x}$ and $\mathbf{x}'$? (Discuss)
- If no, what should be the condition on $\mathbf{W}$ to guarantee that?
- If $\mathbf{W}$ is null (all elements zero), what happens? If $|\mathbf{W}| = 0$, what happens?

Problem - III

We know that the rank of a 3×3 matrix formed by first 9 numbers arranged sequentially is 2.

What is the rank of a 5×5 matrix formed by first 25 numbers arranged sequentially?

Problem - IV

A certain test for disease is known to have True positive of 0.6 and False Positive of 0.1.
A population of 100 people (where 60 of them are infected) undergoes this test.

What could be the confusion matrix?

(a) $\begin{bmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix}$ (b) $\begin{bmatrix} 0.6 & 0.4 \\ 0.9 & 0.1 \end{bmatrix}$ (c) $\begin{bmatrix} 0.6 & 0.2 \\ 0.1 & 0.3 \end{bmatrix}$ (d) $\begin{bmatrix} 0.58 & 0.42 \\ 0.15 & 0.85 \end{bmatrix}$
(e) None of the above

Problem - V

Q: Let us consider that FN rate of TEST-I is 10% and the FP of rate of TEST-II is 50%.
An SMAI student gave a recommendation to the Govt that Every Person should be tested three times (say in a day) and majority label should be assigned.

- Does this make sense? Is this student, technically sound?