# Generative Learning

# and

# Language

**Prepared by: Joseph Bakarji**

# Spam VS. Not Spam

https://www.kaggle.com/datasets/imgowthamg/email-spam-and-non-spam-datasets

| Category | Message |
|---|---|
| ham | Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat... |
| ham | Ok lar... Joking wif u oni... |
| spam | Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's |
| ham | U dun say so early hor... U c already then say... |
| ham | Nah I don't think he goes to usf, he lives around here though |
| spam | FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv |
| ham | Even my brother is not like to speak with me. They treat me like aids patent. |
| ham | As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune |
| spam | WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only. |
| spam | Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030 |
| ham | I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today. |
| spam | SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info |
| spam | URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18 |
| ham | I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times. |
| ham | I HAVE A DATE ON SUNDAY WITH WILL!! |
| spam | XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL |
| ham | Oh k...i'm watching here:) |

B **bar.elli@ajomacm.com** Junk - Exchange August 6, 2024 at 8:50 PM
Dear J. Bakarji: Call for Papers/Abstracts & Editorial Board Member/Reviewer
To: joseph.bakarji@aub.edu.lb

To unsubscribe from our newsletter, please click here.

## Invite You to Send an Article/Abstract

Dear J. Bakarji,

Hope you're doing well.

More than 300 high-quality scientific and scholarly journals are issued by our publishing house, which cover a wide range of disciplines. They are aimed to offer good platforms for researchers to exchange ideas on all fields.

**Invitation to Submit Your Research Articles/Abstracts**

On account of the knowledge shown in your previous publication *"Distilling data into code"*, your contributions of any unpublished research works in your specialized/interested field are welcome.

Abstract/Article Submission Channel:
**http://www.historyrj.org/gs/RYYOW**

**Why Choose the Journal?**

- The articles in these journals are all open access, i.e., freely and publicly accessible online.

**Loren Powlowski - Associate Editor** Junk - Exchange August 3, 2024 at 5:46 PM
Support Our Authors: Explore Publications & Offer Encouragement for Future Work!
To: Bakarji J

**Dear Dr. Bakarji J,**

I trust this email finds you well. I'm reaching out to follow up on an earlier mail I sent this week, for which I haven't received a response.

We greatly admire your dedication to advancing scientific research. Your work **"Distilling data into code"** has been noted by one of our esteemed authors for its insightful contributions to the field.

With that in mind, we would like to extend an invitation with you to share your latest findings and updates in our Upcoming Special Issue on **"Distillation"**. This platform give an excellent Place to expand the reach of your work and foster greater impact within the scientific community.

You're encouraged to submit any recent/unpublished or follow-up studies through our email: submissions@jscimedcentral.com

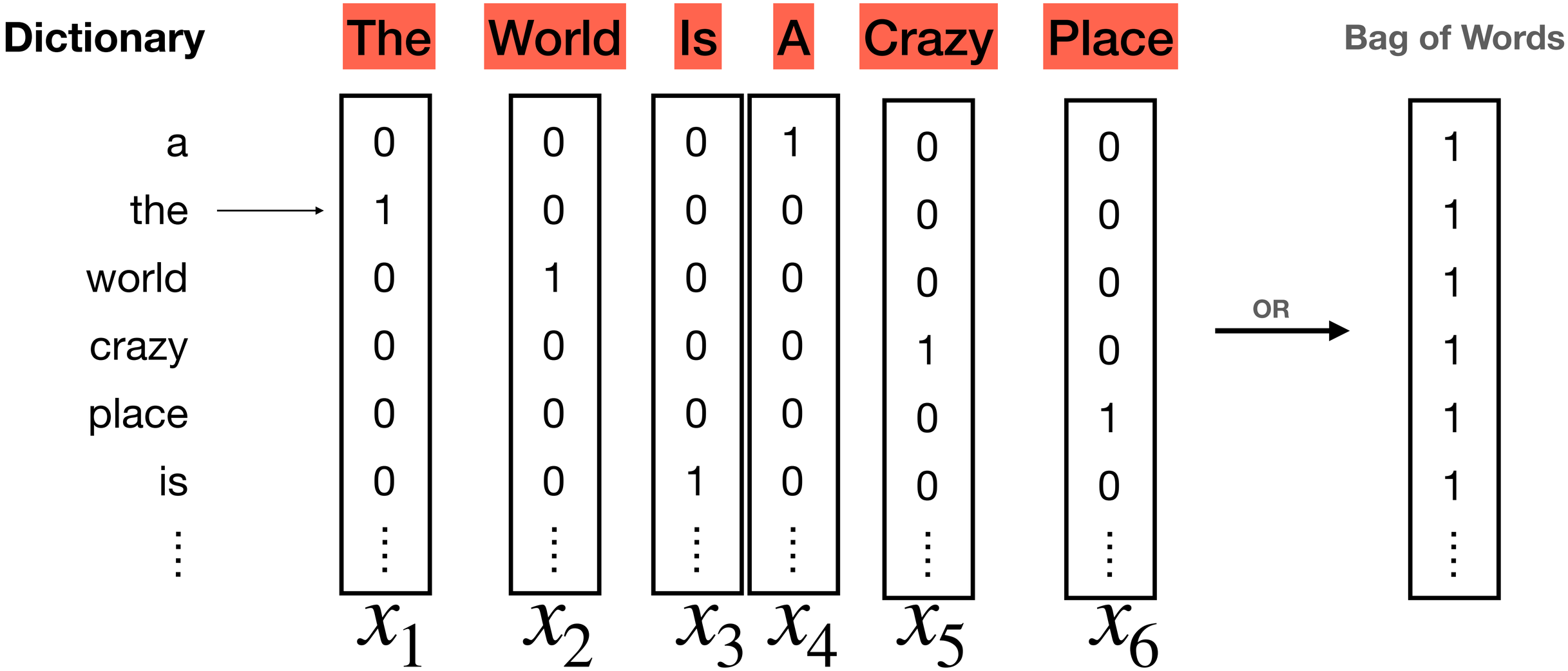
Submission Timeline: on or before **Aug 20th, 2024**

We look forward to the possibility of featuring your work in our publication.

Warm regards,
*Loren Powlowski*
*Associate Editor*
**Chemical Engineering and Process Techniques**

# Language

**One-hot** word representation



# Bag of Words

**What if we define $x_i$ to be a binary feature. Either $x_i = 1$ or $x_i = 0$**

| Email # | x1: free | x2: win | x3: money | … | x4: project | x5: meeting | x6: report | y (label) |
|---|---|---|---|---|---|---|---|---|
| **1** | 1 | 1 | 1 | … | 0 | 0 | 0 | **spam** |
| **2** | 0 | 0 | 0 | … | 1 | 1 | 1 | **not spam** |
| **3** | 1 | 0 | 1 | … | 0 | 0 | 0 | **spam** |
| **4** | 0 | 0 | 0 | … | 1 | 0 | 1 | **not spam** |
| **5** | 1 | 1 | 0 | … | 0 | 1 | 0 | **spam** |
| **6** | 1 | 1 | 0 | … | 1 | 1 | 1 | **not spam** |

$$P(x_i = 1)?$$

# What are the relevant probabilities?

$P(x_i \mid y)$ ⟶

$P(\text{free} = 1 \mid \text{spam}) = ??$

$P(\text{free} = 1 \mid \text{not spam}) = ??$

| Email # | x1: free | x2: win | x3: money | … | x4: project | x5: meetin | x6: report | y (label) |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | … | 0 | 0 | 0 | spam |
| 2 | 0 | 0 | 0 | … | 1 | 1 | 1 | not spam |
| 3 | 1 | 0 | 1 | … | 0 | 0 | 0 | spam |
| 4 | 0 | 0 | 0 | … | 1 | 0 | 1 | not spam |
| 5 | 1 | 1 | 0 | … | 0 | 1 | 0 | spam |
| 6 | 1 | 1 | 0 | … | 1 | 1 | 1 | not spam |

| word | P(word \| spam) | P(word \| not spam) |
|---|---|---|
| Free → | 3/3 | 1/3 |
| Win → | 2/3 | 1/3 |
| Money → | 2/3 | 0/3 |
| Project → | 3/3 | 0/3 |
| Meeting → | 1/3 | 2/3 |
| Report → | 0/3 | 3/3 |

# What are the relevant probabilities?

$$P(x_i \mid y) \longrightarrow \begin{array}{l} P(\text{free} \mid \text{spam}) = 1 \\ P(\text{free} \mid \text{not spam}) = 0 \end{array}$$

$$P(x_1 \mid y)P(x_2 \mid y) \ldots P(x_n \mid y) = \prod_{j=1}^{n} P(x_i \mid y)$$

$$= P(\text{free} = 1 \mid \text{spam})\ P(\text{win} = 1 \mid \text{spam}) \ldots P(\text{report} \mid \text{spam})$$

$$\approx P(x_1, x_2, \ldots, x_n \mid y)$$

??

# What are the relevant probabilities?

$$\approx P(x_1, x_2, \ldots, x_n \mid y) \xrightarrow{??} P(y \mid x_1, x_2, \ldots, x_n)$$

**Probability of spam given words**

**Bayes Rule**

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n \mid y)P(y)}{P(x_1, x_2, \ldots, x_n)} = \frac{P(x \mid y)P(y)}{P(x)}$$

**For a given dataset** $\{(x^{(i)}, y^{(i)})\}_{i=1}^{M}$

$$P(y^{(i)} \mid x^{(i)}) = \frac{P(x^{(i)} \mid y^{(i)})P(y^{(i)})}{P(x^{(i)})}$$

$$P(y^{(i)} = 1 \mid x^{(i)}) = \frac{P(x^{(i)} \mid y^{(i)} = 1)P(y^{(i)} = 1)}{P(x^{(i)})}$$

$$P(y^{(i)} = 1 \mid x^{(i)}) = \frac{\left[\prod_{j=1}^{n} P(x_j^{(i)} \mid y^{(i)} = 1)\right] P(y^{(i)} = 1)}{\left[\prod_{j=1}^{n} P(x_j^{(i)} \mid y^{(i)} = 1)\right] P(y^{(i)} = 1) + \left[\prod_{j=1}^{n} P(x_j^{(i)} \mid y^{(i)} = 0)\right] P(y^{(i)} = 0)}$$

**For a given dataset** $\{(x^{(i)}, y^{(i)})\}_{i=1}^{M}$

$$P(x_j^{(i)} | y^{(i)} = 1) = \frac{\sum_{i=1}^{M} \mathbf{1}\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{M} \mathbf{1}\{y^{(i)} = 1\}}$$

$$P(x_j^{(i)} | y^{(i)} = 0) = \frac{\sum_{i=1}^{M} \mathbf{1}\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{M} \mathbf{1}\{y^{(i)} = 0\}}$$

$$P(y^{(i)} = 1) = \frac{\sum_{i=1}^{M} \mathbf{1}\{y^{(i)} = 1\}}{M}$$

**For a given dataset** $\{(x^{(i)}, y^{(i)})\}_{i=1}^{M}$

$$P(y^{(i)} = 1 \,|\, x^{(i)}) =$$

$$\frac{\left[\prod_{j=1}^{M} P(x_j^{(i)} \,|\, y^{(i)} = 1)\right] P(y^{(i)} = 1)}{\left[\prod_{j=1}^{M} P(x_j^{(i)} \,|\, y^{(i)} = 1)\right] P(y^{(i)} = 1) + \left[\prod_{j=1}^{M} P(x_j^{(i)} \,|\, y^{(i)} = 0)\right] P(y^{(i)} = 0)}$$

# Auto-Regressive Language Models

## N-Gram Model

Given a sequence of words (tokens) $x_i$: $x_1, x_2, x_3, x_4, x_5, x_6 \ldots$

The joint probability of the entire text can be approximated (using the chain rule) as the product of probabilities of finding the next word $x_i$ given the $k$ previous words

$$P(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N-k} P(x_i \mid x_{i-k+1}, \ldots, x_{i-1})$$

Given a corpus, this can be approximated as:

$$P(x_i \mid x_{i-k+1}, \ldots, x_{i-1}) = \frac{\text{Count}(x_{i-k+1}, \ldots, x_{i-1}, x_i)}{\text{Count}(x_{i-k+1}, \ldots, x_{i-1})},$$

Example:

$$P(apple \mid I, ate, an) = \frac{\text{Count}(I, ate, an, apple)}{\text{Count}(I, ate, an)},$$

# Large Language Models

- **A Large Language Model** is also a probabilistic model for $P(x_1, x_2, \ldots, x_N)$
- **A model approximates** $P(x_i \mid x_{i-k+1}, \ldots, x_{i-1})$ which is $V$ dimensional instead of $V^N$ dimensional, $V$ is the size of the dictionary/embedding ($x \in \mathbb{R}^V$)