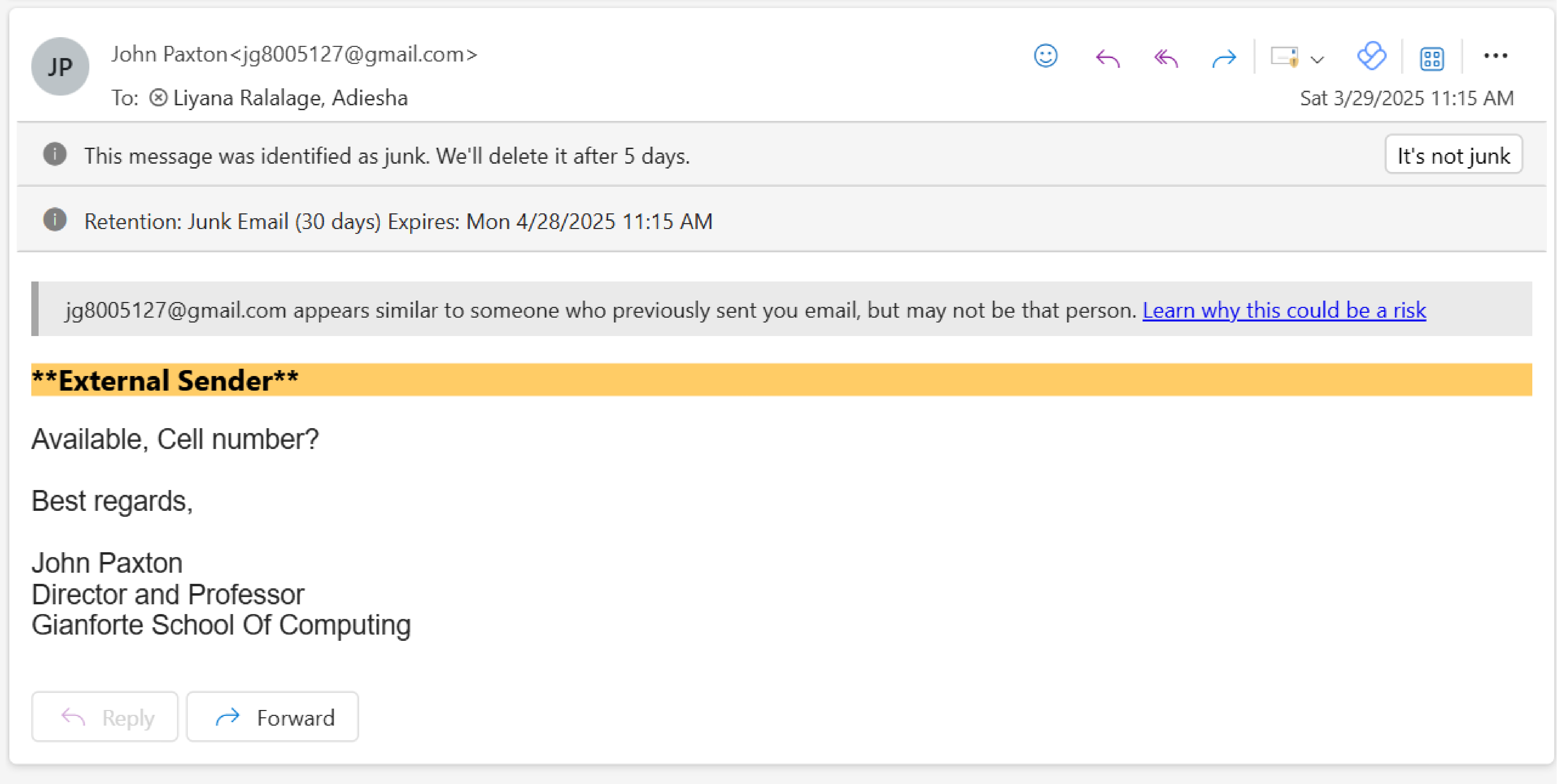
The picture can't be displayed.

Introduction to Text Mining

CSCI 347

Adiesha Liyana Ralalage

Is this spam?



JP John Paxton<jg8005127@gmail.com>

To: ⊗ Liyana Ralalage, Adiesha

Sat 3/29/2025 11:15 AM

This message was identified as junk. We'll delete it after 5 days.

Retention: Junk Email (30 days) Expires: Mon 4/28/2025 11:15 AM

jg8005127@gmail.com appears similar to someone who previously sent you email, but may not be that person. Learn why this could be a risk

External Sender

Available, Cell number?

Best regards,

John Paxton
Director and Professor
Gianforte School Of Computing

The picture can't be displayed.

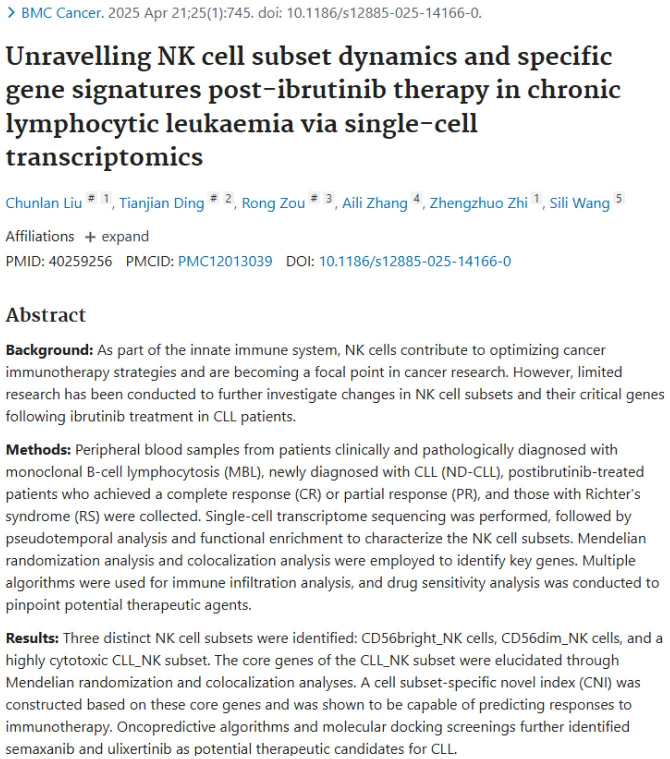
Positive or negative movie review?

- 👎 Unbelievably disappointing
- 👍 Full of zany characters and richly applied satire, and some great plot twists.
- 👍 This is the greatest screwball comedy ever filmed
- 👎 This four-hour endurance test of a movie is a jumble of CGI battles and endless overuse slow motion.
- 👍 Zack Snyder's Justice League is arguably a more cohesive, consistent, and emotionally compelling movie than the 2017's version, despite most of the original Justice League being present in the Snyder Cut.

The picture can't be displayed.

What is the subject of this article?

Medline article



› BMC Cancer. 2025 Apr 21;25(1):745. doi: 10.1186/s12885-025-14166-0.

Unravelling NK cell subset dynamics and specific gene signatures post-ibrutinib therapy in chronic lymphocytic leukaemia via single-cell transcriptomics

Chunlan Liu [# 1], Tianjian Ding [# 2], Rong Zou [# 3], Aili Zhang [4], Zhengzhuo Zhi [1], Sili Wang [5]

Affiliations + expand

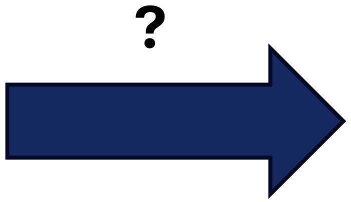
PMID: 40259256 PMCID: PMC12013039 DOI: 10.1186/s12885-025-14166-0

Abstract

Background: As part of the innate immune system, NK cells contribute to optimizing cancer immunotherapy strategies and are becoming a focal point in cancer research. However, limited research has been conducted to further investigate changes in NK cell subsets and their critical genes following ibrutinib treatment in CLL patients.

Methods: Peripheral blood samples from patients clinically and pathologically diagnosed with monoclonal B-cell lymphocytosis (MBL), newly diagnosed with CLL (ND-CLL), postibrutinib-treated patients who achieved a complete response (CR) or partial response (PR), and those with Richter's syndrome (RS) were collected. Single-cell transcriptome sequencing was performed, followed by pseudotemporal analysis and functional enrichment to characterize the NK cell subsets. Mendelian randomization analysis and colocalization analysis were employed to identify key genes. Multiple algorithms were used for immune infiltration analysis, and drug sensitivity analysis was conducted to pinpoint potential therapeutic agents.

Results: Three distinct NK cell subsets were identified: CD56bright_NK cells, CD56dim_NK cells, and a highly cytotoxic CLL_NK subset. The core genes of the CLL_NK subset were elucidated through Mendelian randomization and colocalization analyses. A cell subset-specific novel index (CNI) was constructed based on these core genes and was shown to be capable of predicting responses to immunotherapy. Oncopredictive algorithms and molecular docking screenings further identified semaxanib and ulixertinib as potential therapeutic candidates for CLL.



?

Mesh subject hierarchy

- Anatomy [A]
- Organisms [B]
- Diseases [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]
- Psychiatry and Psychology [F]
- Phenomena and Processes [G]
- Disciplines and Occupations [H]
- Anthropology, Education, Sociology, and Social Phenomena [I]
- ....

The picture can't be displayed.

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Language Identification
- Sentiment analysis

Text Classification: definition

- Input:
  - A document d
  - A fixed set of classes $C = \{c_1, c_2, \dots, c_k\}$
- Output
  - A predicted class $c \in C$

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND "have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

The picture can't be displayed.

Classification Methods:

Any kind of classifier

- Naïve Bayes
- k-Nearest Neighbor
- Logistic regression
- Support-vector machines
- …

Naïve Bayes for Document classification

C_i is the ith class

$$P(C_i|d) = \frac{P(d|C_i) \cdot P(C_i)}{P(d)}$$

d is the document

$$predicted\ class\ = argmax_{C_i}\{P(C_i|d)\} = argmax_{C_i}\{P(d|C_i) \cdot P(C_i)\}$$

Naïve Bayes for Document classification

Each document d is represented as a set of attributes $X_1, X_2, \dots, X_n$

$$P(C_i|d) = \frac{P(d|C_i) \cdot P(C_i)}{P(d)}$$

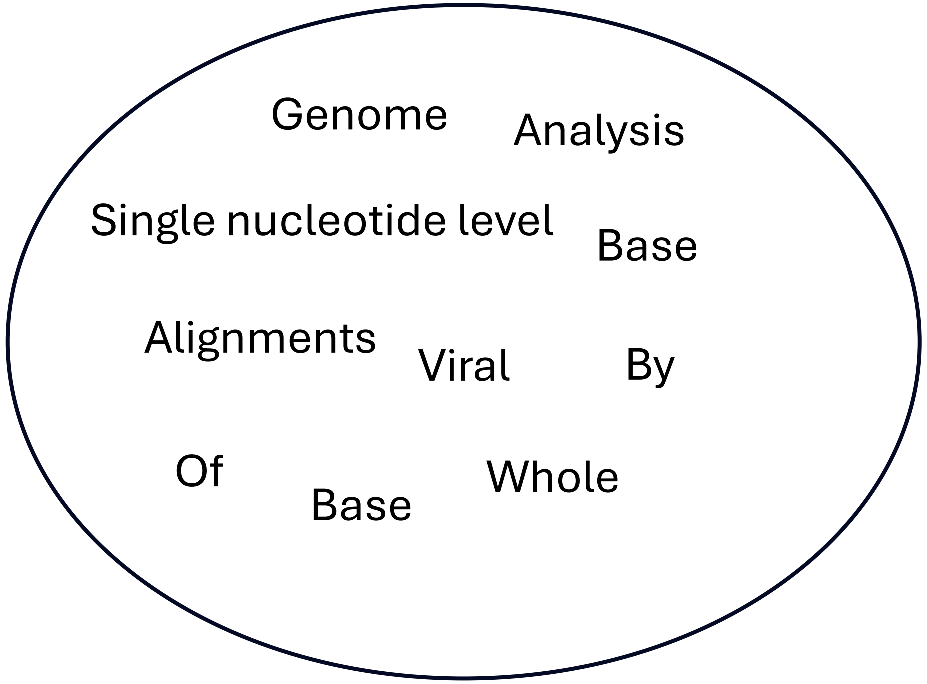
$$\begin{aligned} predicted\ class\ &= argmax_{C_i}\{P(C_i|d)\} = argmax_{C_i}\{P(d|C_i) \cdot P(C_i)\} \\ &= argmax_{C_i}\{P(X_1, X_2, \dots, X_n|C_i)P(C_i)\} \\ &= argmax_{C_i}\{P(X_1|C_i)P(X_2|C_i) \cdots P(X_n|C_i)P(C_i)\} \end{aligned}$$

Naïve Bayes for Document classification

Each document d is represented as a set of attributes $X_1, X_2, \dots, X_n$

We can represent a document as a "bag of words"

d = "Base-By-Base: Single nucleotide-level analysis of whole viral genome alignments"



Genome

Analysis

Single nucleotide level

Base

Alignments

Viral

By

Of

Base

Whole

$$P(C_i|d) = \frac{P(d|C_i) \cdot P(C_i)}{P(d)}$$

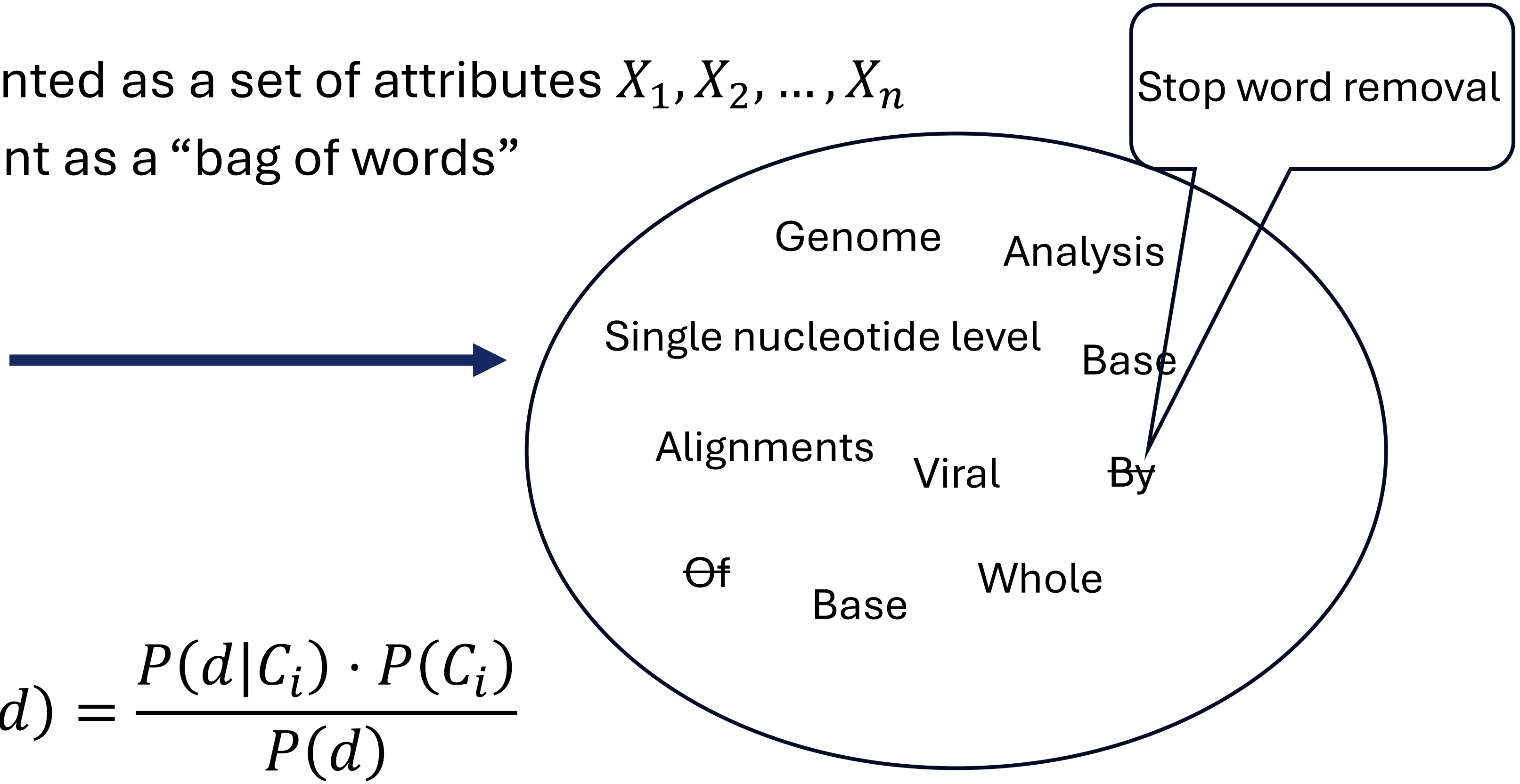
$$\begin{aligned} predicted\ class\ &= argmax_{C_i}\{P(C_i|d)\} = argmax_{C_i}\{P(d|C_i) \cdot P(C_i)\} \\ &= argmax_{C_i}\{P(X_1, X_2, \dots, X_n|C_i)P(C_i)\} \\ &= argmax_{C_i}\{P(X_1|C_i)P(X_2|C_i) \cdots P(X_n|C_i)P(C_i)\} \end{aligned}$$

Naïve Bayes for Document classification

Each document d is represented as a set of attributes $X_1, X_2, \dots, X_n$

We can represent a document as a "bag of words"

d = "Base-By-Base: Single nucleotide-level analysis of whole viral genome alignments"



Stop word removal

Genome

Analysis

Single nucleotide level

Base

Alignments

Viral

~~By~~

~~Of~~

Base

Whole

$$P(C_i|d) = \frac{P(d|C_i) \cdot P(C_i)}{P(d)}$$

$$\begin{aligned} predicted\ class\ &= argmax_{C_i}\{P(C_i|d)\} = argmax_{C_i}\{P(d|C_i) \cdot P(C_i)\} \\ &= argmax_{C_i}\{P(X_1, X_2, \dots, X_n|C_i)P(C_i)\} \\ &= argmax_{C_i}\{P(X_1, X_2, \dots, X_n|C_i)P(C_i)\} \end{aligned}$$

Why do we remove stop words?

- Stop words are very common words in a language — like:
  - "the", "is", "at", "which", "and" "a", "of", "in", etc.
  - Very high frequency
- They carry little meaning
  - Stop words don’t help distinguish between different categories of documents.
  - Example: "the cat" vs "the dog" → "the" adds no classification value.
- Reduce noise
- Reduce dimensionality

Naïve Bayes for Document classification

$$P(C_i|d) = \frac{P(d|C_i) \cdot P(C_i)}{P(d)}$$

Document	Class (Topic)
d_1 = “The Virus That Changed My World”	Mitigation
d_2 = “Base-By-Base: Single nucleotide-level analysis of whole viral genome alignments”	Vaccine
$d_3 =$ “The immediate effects of the severe acute respiratory syndrome (SARS) epidemic on childbirth in Taiwan”	Rist factors
$d_4 =$ “Persistence of lung inflammation and lung cytokines with high-resolution CT abnormalities during recovery from SARS”	Risk factors
$d_5 =$ Date of origin of the SARS coronavirus strains”	Mitigation
$d_6 =$ “Design of Wide-Spectrum Inhibitors Targeting Coronavirus Main Proteases”	Vaccine
$d_7 =$ “A systematic review of the effectiveness of antimicrobial rinse-free hand sanitizers for prevention of illness-related absenteeism in elementary school”	Mitigation
$d_8 =$ “Globalization and risks to health”	Risk factors

$$predicted\ class\ = argmax_{C_i}\{P(C_i|d)\} = argmax_{C_i}\{P(X_1, X_2, \dots, X_n|C_i)P(C_i)\}$$

Naïve Bayes for Document classification

$$P(C_i|d) = \frac{P(d|C_i) \cdot P(C_i)}{P(d)}$$

Document	Class (Topic)
d_1 = “~~The~~ Virus ~~That~~ Changed My World”	Mitigation
d_2 = “Base-~~By~~-Base: Single nucleotide-level analysis ~~of~~ whole viral genome alignments”	Vaccine
d_3 = “~~The~~ immediate effects ~~of the~~ severe acute respiratory syndrome (SARS) epidemic ~~on~~ childbirth ~~in~~ Taiwan”	Rist factors
d_4 = “Persistence ~~of~~ lung inflammation ~~and~~ lung cytokines ~~with~~ high-resolution CT abnormalities during recovery ~~from~~ SARS”	Risk factors
d_5 = Date ~~of~~ origin ~~of the~~ SARS coronavirus strains”	Mitigation
d_6 = “Design ~~of~~ Wide-Spectrum Inhibitors Targeting Coronavirus Main Proteases”	Vaccine
d_7 = “~~A~~ systematic review ~~of the~~ effectiveness of antimicrobial rinse-free hand sanitizers ~~for~~ prevention ~~of~~ illness-related absenteeism ~~in~~ elementary school”	Mitigation
d_8 = “Globalization ~~and~~ risks ~~to~~ health”	Risk factors

$$predicted\ class\ = argmax_{C_i}\{P(C_i|d)\} = argmax_{C_i}\{P(X_1, X_2, \ldots, X_n|C_i)P(C_i)\}$$

Naïve Bayes for Document classification

$$P(C_i|d) = \frac{P(d|C_i) \cdot P(C_i)}{P(d)}$$

Document	X_1: “SARS”	X_2: “Genome”	X_3: “Effectiveness”	$\cdots$	X_m: “inhibitors”	Class (Topic)
d_1	0	0	0	$\cdots$	0	Vaccine
d_2	0	1	0	$\cdots$	0	Mitigation
d_3	1	0	0	$\cdots$	0	Risk factor
d_4	1	0	0	$\cdots$	0	Risk factor
d_5	1	0	0	$\cdots$	0	Mitigation
d_6	0	0	0	$\cdots$	1	Vaccine
d_7	0	0	1	$\cdots$	0	Mitigation
d_8	0	0	0	$\cdots$	0	Risk factor

$$predicted\ class\ = argmax_{C_i}\{P(C_i|d)\} = argmax_{C_i}\{P(X_1, X_2, \ldots, X_n|C_i)P(C_i)\}$$

Naïve Bayes for Document classification $P(C_i|d) = \frac{P(d|C_i) \cdot P(C_i)}{P(d)}$

Document	X_1: “SARS”	X_2: “Genome”	X_3: “Effectiveness”	$\cdots$	X_m: “inhibitors”	Class (Topic)
d_1	0	0	0	$\cdots$	0	Vaccine
d_2	0	1	0	$\cdots$	0	Mitigation
d_3	1	0	0	$\cdots$	0	Risk factor
d_4	1	0	0	$\cdots$	0	Risk factor
d_5	1	0	0	$\cdots$	0	Mitigation
d_6	0	0	0	$\cdots$	1	Vaccine
d_7	0	0	1	$\cdots$	0	Mitigation
d_8	0	0	0	$\cdots$	0	Risk factor

Can be binary (presence/absence) or counts/frequencies of words

$$predicted\ class\ = argmax_{C_i}\{P(C_i|d)\} = argmax_{C_i}\{P(X_1, X_2, \ldots, X_n|C_i)P(C_i)\}$$

The picture can't be displayed.

Multinomial Naïve Bayes

When considering the counts of each word we can use multinomial naïve bayes

- Assumptions
- Bag of Words assumption: Assume position doesn't matter
- Conditional Independence: Assume the feature probabilities $P\left(x_i \middle| c_j\right)$ are independent given the class c.

$$P(x_1, \ldots, x_n|c) = P(x_1|c) \cdot P(x_2|c) \cdots P(x_n|c)$$

Multinomial Naïve Bayes classifier

When considering the counts of each word we can use multinomial naïve bayes

$$c_{MAP} = argmax_{c \in C}\, \{P(x_1, x_2, \dots, x_n | c) P(c)\}$$

$$c_{MAP} = argmax_{c \in C}\, \left\{ P(c) \prod_{x \in \mathrm{X}} P(x|c) \right\}$$

- MAP is “maximum a posteriori” = most likely class

Multinomial Naïve Bayes classifier

When considering the counts of each word we can use multinomial naïve bayes

$$positions \leftarrow all\ word\ positions\ in\ test\ document$$

$$c_{MAP} = argmax_{c \in C} \left\{ P(c) \prod_{x \in positions} P(x|c) \right\}$$

- MAP is “maximum a posteriori” = most likely class

Multinomial Naïve Bayes classifier

There is a problem with this method

$$c_{MAP} = argmax_{c \in C} \left\{ P(c) \prod_{x \in positions} P(x|c) \right\}$$

- When multiplying probabilities, computation can result in a floating-point underflow!
- So, we use logs, because $\log ab = \log a + \log b$
- We will sum up the probabilities instead of multiplying.

The picture can't be displayed.

Multinomial Naïve Bayes classifier

Instead of $c_{MAP} = argmax_{c \in C} \left\{ P(c) \prod_{x \in positions} P(x|c) \right\}$

We calculate $C_{NB} = argmax_{c \in C} \left[\log P\left(c_j\right) + \sum_{i \in positions} \log P\left(x_i | c_j\right) \right]$

The picture can't be displayed.

Multinomial Naïve Bayes classifier

How to estimate the parameters

Maximum likelihood and prior parameter estimates

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

$\hat{P}(w_i|c_j) = \frac{count(w_i,c_j)}{\sum_{w \in V} count(w,c_j)}$ fraction of time word w_i appears among all words in documents of topic c_j

The picture can't be displayed.

Multinomial Naïve Bayes classifier

Problems with estimating maximum likelihood

What if we have seen no training documents with the word **awesome** and classified in the topic **positive** (thumbs-up)?

$$\hat{P}(\text{"}awesome\text{"}|postive) = \frac{count(\text{"}fantastic\text{"}, postive)}{\sum_{w \in V} count(w, postive)} = 0$$

Zero probabilities cannot be conditioned away, no matter the other

Evidence!

$$c_{MAP} = argmax_{c \in C} \left\{ P(c) \prod_{x \in positions} P(x|c) \right\}$$

The picture can't be displayed.

Laplace (add 1) smoothing for naïve bayes

Problems with estimating maximum likelihood

What if we have seen no training documents with the word **awesome** and classified in the topic **positive** (thumbs-up)?

$$\hat{P}\left(w_i|c_j\right) = \frac{count\left(w_i,c_j\right)+1}{\sum_{w\in V}\left(count\left(w,c_j\right)+1\right)} = \frac{count\left(w_i,c_j\right)+1}{\sum_{w\in V}\left(count\left(w,c_j\right)\right)+|V|}$$