به نام خدا

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

آزمونک شماره دو - پردازش زبان طبیعی

سوال ۱) اسناد زیر را در اختیار داریم:

Document	Text	Class
1	I loved the movie	+
2	I hated the movie	-
3	a great movie, really good	+
4	poor acting	-
5	not a good play	-
6	great acting, not a bad movie	+
7	I loved the poor play	?
8	I hated the acting	?

با استفاده از طبقه‌بند Naïve Bayes، مشخص کنید که اسناد ۷ و ۸ در کدام یک از کلاس‌های مثبت/ منفی قرار می‌گیرند. (از روش add-1 smoothing در محاسبات خود استفاده کنید.) (۳۰ نمره)

Vocabulary = {I, loved, the, movie, hated, a, great, really, good, poor, acting, not, bad, play}

$$P(w_k \mid +) = \frac{n_k+1}{n + |vocabulary|}$$

$$P(+) = \frac{Number\ of\ positive\ documents}{Total\ number\ of\ documents} = \frac{3}{6} = 0.5$$

$$P(-) = \frac{Number\ of\ negative\ documents}{Total\ number\ of\ documents} = \frac{3}{6} = 0.5$$

"I loved the poor play"

P (I | +) = $\frac{1+1}{15+14} = \frac{2}{29} = 0.0689$

P (loved | +) = $\frac{1+1}{15+14} = \frac{2}{29} = 0.0689$

P (the | +) = $\frac{1+1}{15+14} = \frac{2}{29} = 0.0689$

P (poor | +) = $\frac{0+1}{15+14} = \frac{1}{29} = 0.0344$

P (play | +) = $\frac{0+1}{15+14} = \frac{1}{29} = 0.0344$

P (I | -) = $\frac{1+1}{10+14} = \frac{1}{12} = 0.0833$

P (loved | -) = $\frac{0+1}{10+14} = \frac{1}{24} = 0.0416$

P (the | -) = $\frac{1+1}{10+14} = \frac{1}{12} = 0.0833$

P (poor | -) = $\frac{1+1}{10+14} = \frac{1}{12} = 0.0833$

P (play | -) = $\frac{1+1}{10+14} = \frac{1}{12} = 0.0833$

If v_j = +,

P(+) P (I | +) P (loved | +) P (the | +) P (poor | +) P (play | +) = $0.5 * (\frac{2}{29})^3 * (\frac{1}{29})^2 = 1.950 * 10^{-7}$

If v_j = -

P(-) P (I | -) P (loved | -) P (the | -) P (poor | -) P (play | -) = $0.5 * (\frac{1}{12})^4 * (\frac{1}{24}) = 1.005 * 10^{-6}$

So, the posterior probability for class "-" is greater than class "+". Therefore, according to the Naive Bayes classifier, Document 7 would be classified as negative ("-").

"I hated the acting"

P (I | +) = $\frac{1+1}{15+14} = \frac{2}{29} = 0.0689$

P (hated | +) = $\frac{0+1}{15+14} = \frac{1}{29} = 0.0344$

P (the | +) = $\frac{1+1}{15+14} = \frac{2}{29} = 0.0689$

P (acting | +) = $\frac{0+1}{15+14} = \frac{1}{29} = 0.0344$

P (I | -) = $\frac{1+1}{10+14} = \frac{1}{12} = 0.0833$

P (hated | -) = $\frac{1+1}{10+14} = \frac{1}{12} = 0.0833$

P (the | -) = $\frac{1+1}{10+14} = \frac{1}{12} = 0.0833$

P (acting | -) = $\frac{1+1}{10+14} = \frac{1}{12} = 0.0833$

If v_j = +,

P(+) P (I | +) P (loved | +) P (the | +) P (poor | +) P (play | +) = $0.5 * (\frac{2}{29})^2 * (\frac{1}{29})^2 = 2.827 * 10^{-6}$

If v_j = -

P(-) P (I | -) P (loved | -) P (the | -) P (poor | -) P (play | -) = $0.5 * (\frac{1}{12}) = 2.411 * 10^{-5}$

Since the posterior probability for class "-" is greater than for class "+", according to the Naive Bayes classifier, Document 8 would be classified as negative ("-").

سوال ۲) به سوالات زیر پاسخ دهید. (۴۰ نمره).

الف) مدل Skip-gram را در نظر بگیرید.

$$P(\text{context} = y|\text{word} = x) = \frac{\exp(\mathbf{v}_x \cdot \mathbf{c}_y)}{\sum_{y'} \exp(\mathbf{v}_x \cdot \mathbf{c}_{y'})}$$

پیچیدگی محاسباتی برای این احتمال به چه صورت می‌باشد؟

$$P(\text{context} = c|\text{word} = w)$$

این پیچیدگی را براساس d (dimensionality of the vectors) و |V| (vocabulary size) بیان کنید.

O(dv). The dot product is linear in d and not quadratic.

ب) تفاوت مدل‌های Skip-gram و CBOW را توضیح دهید.

The main difference between CBOW and Skip-gram lies in their approach to the prediction task: CBOW predicts a word given its context, while Skip-gram predicts the context given a word. This leads to differences in efficiency, with CBOW being generally faster but Skip-gram offering advantages in quality, particularly for rare words.

پ) آیا آموزش یک skip-gram با negative sampling، نسبت به آموزش مدل اصلی skip-gram سریع‌تر است؟ چرا؟

Computing the normalization constant for skip-gram during learning requires normalizing over the vocabulary, whereas skip-gram with negative sampling does not. You can also see this as only reducing the scores of certain sampled contexts rather than all negative contexts.

ت) در استفاده از روش Bag of Word جهت انجام تسک تحلیل احساسات، هنگام برخورد با کلمه not که کل قطبیت را معکوس می‌کند، چه اقداماتی می‌توان انجام داد؟

For handling the word "not" which can reverse the sentiment polarity in a sentence when using the Bag of Words (BoW) method for sentiment analysis, one strategy is to modify the words in the sentence from the word "not" up to the next punctuation mark by appending the suffix "_NOT" to each word, thus creating new terms.

This approach helps to distinguish the negated context in the processed text. By doing so, words that are part of a negated context are treated as distinct tokens compared to their positive counterparts.

سوال ۳) مدلی را آموزش داده‌ایم تا بررسی کند که آیا یک ایمیل دریافت شده Spam است یا خیر. پس از آموزش مدل، آن را بر روی ۵۰۰ داده‌ی جدید (برچسب‌دار) تست کرده‌ایم و نتایج به شکل زیر است.

		True Class	
		Spam	Not Spam
Predicted Class	Spam	70	30
	Not Spam	70	330

الف) مقادیر **Macro Average** و **Micro Average** را برای معیار **F1-Score** با در نظر گرفتن هر دو کلاس محاسبه کنید. (۲۰ نمره).

$$\text{Precision for Spam} = \frac{70}{70+30} = 0.7$$

$$\text{Precision for Not Spam} = \frac{330}{330+70} = 0.825$$

$$\text{Recall for Spam} = \frac{70}{70+70} = 0.5$$

$$\text{Recall for Not Spam} = \frac{330}{330+30} = 0.916$$

Macro Average F1-Score

$$\text{F1-Score for Spam} = \frac{2\,(P\ Spam * R\ Spam)}{P\ Spam + R\ Spam} = \frac{2\,(0.7 * 0.5)}{0.7+0.5} = 0.583$$

$$\text{Fl-Score for Not Spam} = \frac{2\,(P\ Not\ Spam * R\ Not\ Spam)}{P\ Not\ Spam + R\ Not\ Spam} = \frac{2\,(0.825 * 0.916)}{0.825+0.916} = 0.868$$

$$\text{Macro Average F1-Score} = \frac{0.583 + 0.865}{2} = 0.724$$

Micro Average F1-Score

$$\text{Micro Precision} = \frac{Total\ TP}{Total\ TP + Total\ FP} = \frac{70 + 330}{70 + 330 + 30 + 70} = 0.8$$

$$\text{Micro Recall} = \frac{Total\ TP}{Total\ TP + Total\ FN} = \frac{70 + 330}{70 + 330 + 30 + 70} = 0.8$$

$$\text{Micro F1-Score} = 2 * \frac{Micro\ Precision * Micro\ Recall}{Micro\ Precision + Micro\ Recall} = 0.8$$

ب) فرض کنید که دو کاربر داریم که می خواهند از این مدل استفاده کنند:

کاربر A از اینکه در Inbox ایمیلش، Spam ببیند خوشش نمی‌آید. هرچند مشکلی با اینکه قسمت Junk ایمیلش را مدام چک کند تا ایمیل‌هایی را که به اشتباه Spam درنظر گرفته شده‌اند را بررسی کند، ندارد.

در مقابل، **کاربر B** حتی مطلع نیست که قسمتی به نام Junk وجود دارد تا بتواند ایمیل‌هایی را که Spam تشخیص داده شده‌اند، بررسی کند. در نتیجه ترجیح می‌دهد برخی از ایمیل‌های Spam را (اشتباها) در Inbox ببیند، به جای آنکه، بدون اینکه مطلع باشد، ایمیل‌های مهم را از دست بدهد.

حال، به‌نظر شما کدام یک از این کاربران احساس بهتری نسبت به استفاده از این مدل خواهند داشت؟ چرا؟ (۲۰ نمره).

In order to answer this question, let's think about what it means to have high precision and low recall with respect to SPAM and, conversely, what it means to have high recall and low precision with respect to SPAM.

High-precision and low recall with respect to SPAM: whatever the model classifies as SPAM is probably SPAM. However, many emails that are truly SPAM are misclassified as NOT SPAM. The user is likely to see some SPAM messages in his/her inbox, but will never have to go to the "junk" directory to look for genuine messages incorrectly marked as SPAM.

High recall and low precision with respect to SPAM: the model filters all the SPAM emails, but also incorrectly classifies some genuine emails as SPAM. The user will never see SPAM emails in his/her inbox, but will have to periodically check the "junk" directory for genuine emails incorrectly marked as SPAM.

Because the classifier achieves higher precision than recall, ***USER B*** is more likely to be satisfied with the classifier.