Natural Language Processing

Daniel Jeong

October 19th, 2025

Contents

1. What are some key pieces of knowledge/ skills you've learned from the first half of the class?

   Tracing the evolution of NLP has deepened my appreciation for the fundamental linguistic concepts I've been using unconsciously all along. One of the most striking ideas from the first half of the course is the separation of grammar from meaning. The realization that a sentence can be structurally perfect yet entirely nonsensical. This concept, alongside the application of probability through tools like n-grams, has been particularly insightful, even when acknowledging the inherent limitations they face against the infinite nature of our language. Tokenization, especially sub-word tokenization, was fascinating as well. This prompted me to explore how Korean is tokenized, given that its "syllable block" structure and conjugation rules differ so dramatically from English.

2. How does what you've learned so far relate to what you are learning from other classes and/or help you in your academic or non-academic life?

   While doing Lab06, something very intriguing came up. We used word embeddings and calculated cosine similarities to find the closest word to "doctor - man + woman," and "nurse" was the result. Unlike "actress" or "goddess," "doctor" and "nurse" aren't gender-specific nouns. This result, and the concepts we are learning in the course, make me more rigorously scrutinize the biases I have in my words to be more objective and free from irrelevant noise.

3. What were your goals for taking this class? How does what you've learned so far fit in with these goals?

   I looked forward to replicating papers, and the midterm project was exactly what I was looking for.

4. How do you feel you have been performing in this class? What are some things you are proud of? What are some things that can be improved?

   I really like the feeling of learning the "intuition" of NLP. I think some of the missing documentation was a blessing in disguise because it nudged me to form hypotheses and experiment, and that really helped me build intuition and a bias for action.

5. What were some strategies/ techniques you used or discovered that helped you in the first half of the semester?

   Whatever the problem was, just starting to try different things on a smaller scale ultimately led to a roughly correct solution faster than trying to solve it in one go. In general, I really hate the feeling of being wrong or incorrect, so I tend to try to be perfect on the first attempt, but that approach always backfired.

6. Is there anything that you would want to do differently in the second half of the semester to improve what you can get out of this class?

   I would be more "aggresive"/passionante in my questions.

7. Is there something else you would like to add or reflect upon?

   Overall, I really liked working in groups, especially with someone who had a different background or perspective. I especially liked working with Will and Jordan, and the experience demonstrated how great it can be to work with people who have complementary skill sets. I wonder if I was as great of a teammate as I had hoped to be and would love to get feedback from others. I know anonymous feedback can be hurtful or, on the other end of the spectrum, just too "flattering" to be useful, but I would still love to learn how I can improve.

Chapter 1

Examples

1.1 What should I name it?

Notation:

$x \in S$: x is an element of set S
$x \notin S$: x is not an element of S
$\mathbb{Z}$: The set of integers; {... , -2, -1, 0, 1, 2, ...}
$\mathbb{Z}^+$: The set of positive integers; {1, 2, ...}

Definition 1: Divide

If m and n are integers, we say that m divides n, and we write $m|n$, if there exists some $c \in \mathbb{Z}$ such that $n = mc$.
If $n \neq mc$, m does not divide n, and we write $m \nmid n$.

Pitfall 1 Zero **CAN** divide zero

Since $0 = 0 \times c$ is true for any integer c, 0 divides 0.

Pitfall 2 Integer c **CAN** be negative

$a \mid b$ and $b \mid a$ does not mean $a = b$.
e.g. a = 4, b = -4

Definition 2: Greatest Common Divisor (gcd)

If m and n are integers and not both are 0, then the *greatest comon divisor* of m and n is the largest integer that divides both m and n. We denote the gcd of m and n by (m, n) or gcd(m, n).

e.g., $(12, 18) = 6$.
gcd(0, 0) is undefined because every integer can divide 0.

Insight 1 gcd is always positive

e.g., (-45, 27) = 9. (-28, 39) = 1

Theorem 1

For all integers a, b, c, d:

(i) If $a \mid b$ and $b \mid c$ then $a \mid c$

(ii) If $a \mid b$ and $a \mid c$ then for all $x, y \in \mathbb{Z}$, $a \mid (xb + yc)$

(iii) If $a \mid b$ and $c \mid d$ then $ac \mid bd$

Proof: Since $a \mid b$, there exists $k \in \mathbb{Z}$ such that $b = ak$
Since $b \mid c$, there exists $l \in \mathbb{Z}$ that $c = bl$
Then

$$\begin{aligned} c &= bl \\ &= (ak)l \\ &= a(kl) \end{aligned}$$

Since $k, l \in \mathbb{Z}$, $kl \in \mathbb{Z}$.
So since $c = a(kl)$, $a \mid c$. □ 😊

Theorem 2

Suppose a and b are integers that are not both 0, so that (a, b) exists. b
Then for every $n \in \mathbb{Z}$,

$$\begin{aligned} (a, b) &= (a + nb, b) \\ &= (a, b + na) \end{aligned}$$

skipped Proof of Theorem 2

Definition 3: Euclidean Algorithm

The Euclidean Algorithm is a method for finding the greatest common divisor (GCD) of two numbers. It works as follows:

1. Take two positive integers a and b.
2. Divide a by b. If the remainder is 0, b is the GCD.
3. If not, replace a with b, and b with the remainder.
4. Repeat steps 2-3 until the remainder is 0.

The last non-zero remainder is the GCD of the original two numbers.
This algorithm works because the GCD of two numbers is also the GCD of the smaller number and the remainder of the larger number divided by the smaller one.

Lemma 1 Bezout's Lemma

If m and n are integers that are not both zero that (m, n) exists,
then there exist $x, y \in \mathbb{Z}$ such that $xm + yn = (m, n)$

Definition 4: Coprime

We say integers a and b are relatively prime(or "coprime") if (a,b)=1

Question 1

Is the set $x-$axis\{Origin} a closed set

Solution: We have to take its complement and check whether that set is a open set i.e. if it is a union of open balls

Note:-

We will do topology in Normed Linear Space (Mainly $\mathbb{R}^n$ and occasionally $\mathbb{C}^n$)using the language of Metric Space

Claim 1.1.1 Topology

Topology is cool

Example 1.1.1 (Open Set and Close Set)

Open Set:
- ϕ
- $\bigcup_{x \in X} B_r(x)$ (Any $r > 0$ will do)
- $B_r(x)$ is open

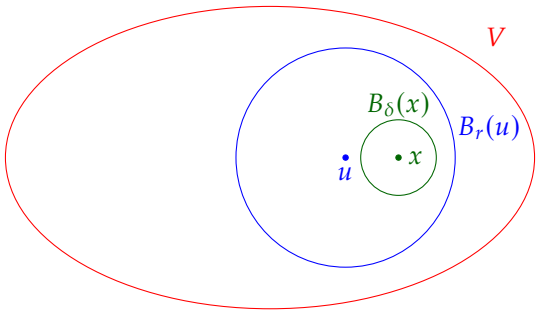
Closed Set:
- $X,\ \phi$
- $\overline{B_r(x)}$

$x-$axis $\cup$ $y-$axis

Theorem 3

If $x \in$ open set V then $\exists\ \delta > 0$ such that $B_\delta(x) \subset V$

Proof: By openness of V, $x \in B_r(u) \subset V$



Given $x \in B_r(u) \subset V$, we want $\delta > 0$ such that $x \in B_\delta(x) \subset B_r(u) \subset V$. Let $d = d(u, x)$. Choose δ such that $d + \delta < r$ (e.g. $\delta < \frac{r-d}{2}$)

If $y \in B_\delta(x)$ we will be done by showing that $d(u, y) < r$ but

$$d(u, y) \leqslant d(u, x) + d(x, y) < d + \delta < r$$

😄

Corollary 1.1.1

By the result of the proof, we can then show...

Lemma 2

Suppose $\vec{v_1}, \ldots, \vec{v_n} \in \mathbb{R}^n$ is subspace of $\mathbb{R}^n$.

Proposition 1.1.1

$1 + 1 = 2$.

1.2 Random Examples

Definition 5: Limit of Sequence in $\mathbb{R}$

Let $\{s_n\}$ be a sequence in $\mathbb{R}$. We say

$$\lim_{n\to\infty} s_n = s$$

where $s \in \mathbb{R}$ if $\forall$ real numbers $\epsilon > 0$ $\exists$ natural number N such that for $n > N$

$$s - \epsilon < s_n < s + \epsilon \text{ i.e. } |s - s_n| < \epsilon$$

Question 2

Is the set $x-$axis$\backslash\{$Origin$\}$ a closed set

Solution: We have to take its complement and check whether that set is a open set i.e. if it is a union of open balls

Note:-

We will do topology in Normed Linear Space (Mainly $\mathbb{R}^n$ and occasionally $\mathbb{C}^n$)using the language of Metric Space

Claim 1.2.1 Topology

Topology is cool

Example 1.2.1 (Open Set and Close Set)

Open Set:
- ϕ
- $\bigcup_{x\in X} B_r(x)$ (Any $r > 0$ will do)
- $B_r(x)$ is open

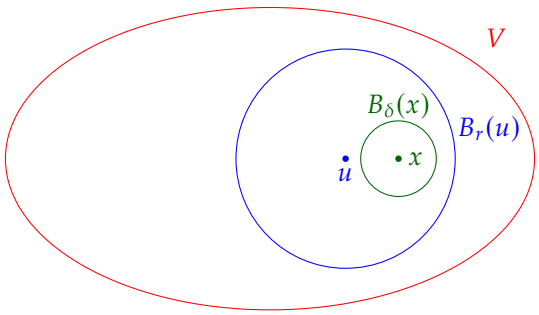
Closed Set:
- X, ϕ
- $\overline{B_r(x)}$

$x-$axis $\cup$ $y-$axis

Theorem 4

If $x \in$ open set V then $\exists$ $\delta > 0$ such that $B_\delta(x) \subset V$

Proof: By openness of V, $x \in B_r(u) \subset V$



Given $x \in B_r(u) \subset V$, we want $\delta > 0$ such that $x \in B_\delta(x) \subset B_r(u) \subset V$. Let $d = d(u,x)$. Choose δ such that $d + \delta < r$ (e.g. $\delta < \frac{r-d}{2}$)

If $y \in B_\delta(x)$ we will be done by showing that $d(u,y) < r$ but

$$d(u,y) \leqslant d(u,x) + d(x,y) < d + \delta < r$$

😄

Corollary 1.2.1

By the result of the proof, we can then show...

Lemma 3

Suppose $\vec{v_1}, \ldots, \vec{v_n} \in \mathbb{R}^n$ is subspace of $\mathbb{R}^n$.

Proposition 1.2.1

$1 + 1 = 2$.

1.3 Random

Definition 6: Normed Linear Space and Norm $\|\cdot\|$

Let V be a vector space over $\mathbb{R}$ (or $\mathbb{C}$). A norm on V is function $\|\cdot\|\ V \to \mathbb{R}_{\geqslant 0}$ satisfying

① $\|x\| = 0 \iff x = 0\ \forall\ x \in V$

② $\|\lambda x\| = |\lambda|\|x\|\ \forall\ \lambda \in \mathbb{R}(\text{or } \mathbb{C}),\ x \in V$

③ $\|x + y\| \leqslant \|x\| + \|y\|\ \forall\ x, y \in V$ (Triangle Inequality/Subadditivity)

And V is called a normed linear space.
• Same definition works with V a vector space over $\mathbb{C}$ (again $\|\cdot\| \to \mathbb{R}_{\geqslant 0}$) where ② becomes $\|\lambda x\| = |\lambda|\|x\|$ $\forall\ \lambda \in \mathbb{C}$, $x \in V$, where for $\lambda = a + ib$, $|\lambda| = \sqrt{a^2 + b^2}$

Example 1.3.1 ($p-$Norm)

$V = \mathbb{R}^m$, $p \in \mathbb{R}_{\geqslant 0}$. Define for $x = (x_1, x_2, \cdots, x_m) \in \mathbb{R}^m$

$$\|x\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_m|^p\right)^{\frac{1}{p}}$$

(In school $p = 2$)

Special Case $p = 1$: $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_m|$ is clearly a norm by usual triangle inequality.
Special Case $p \to \infty$ ($\mathbb{R}^m$ with $\|\cdot\|_\infty$): $\|x\|_\infty = \max\{|x_1|, |x_2|, \cdots, |x_m|\}$
For $m = 1$ these $p-$norms are nothing but $|x|$. Now exercise

Question 3

Prove that triangle inequality is true if $p \geqslant 1$ for $p-$norms. (What goes wrong for $p < 1$?)

Solution: **For Property ③ for norm-2**

When field is $\mathbb{R}$:

We have to show

$$\sum_i (x_i + y_i)^2 \leqslant \left(\sqrt{\sum_i x_i^2} + \sqrt{\sum_i y_i^2} \right)^2$$

$$\implies \sum_i (x_i^2 + 2x_i y_i + y_i^2) \leqslant \sum_i x_i^2 + 2\sqrt{\left[\sum_i x_i^2\right]\left[\sum_i y_i^2\right]} + \sum_i y_i^2$$

$$\implies \left[\sum_i x_i y_i\right]^2 \leqslant \left[\sum_i x_i^2\right]\left[\sum_i y_i^2\right]$$

So in other words prove $\langle x, y\rangle^2 \leqslant \langle x, x\rangle\langle y, y\rangle$ where

$$\langle x, y\rangle = \sum_i x_i y_i$$

Note:-

- $\|x\|^2 = \langle x, x\rangle$
- $\langle x, y\rangle = \langle y, x\rangle$
- $\langle \cdot, \cdot \rangle$ is $\mathbb{R}-$linear in each slot i.e.

  $\langle rx + x', y\rangle = r\langle x, y\rangle + \langle x', y\rangle$ and similarly for second slot

  Here in $\langle x, y\rangle$ x is in first slot and y is in second slot.

Now the statement is just the Cauchy-Schwartz Inequality. For proof

$$\langle x, y\rangle^2 \leqslant \langle x, x\rangle\langle y, y\rangle$$

expand everything of $\langle x - \lambda y, x - \lambda y\rangle$ which is going to give a quadratic equation in variable λ

$$\begin{aligned}\langle x - \lambda y, x - \lambda y\rangle &= \langle x, x - \lambda y\rangle - \lambda\langle y, x - \lambda y\rangle \\ &= \langle x, x\rangle - \lambda\langle x, y\rangle - \lambda\langle y, x\rangle + \lambda^2\langle y, y\rangle \\ &= \langle x, x\rangle - 2\lambda\langle x, y\rangle + \lambda^2\langle y, y\rangle\end{aligned}$$

Now unless $x = \lambda y$ we have $\langle x - \lambda y, x - \lambda y\rangle > 0$ Hence the quadratic equation has no root therefore the discriminant is greater than zero.

When field is $\mathbb{C}$:

Modify the definition by

$$\langle x, y\rangle = \sum_i \overline{x_i} y_i$$

Then we still have $\langle x, x\rangle \geqslant 0$

1.4 Algorithms

Algorithm 1: what about

Input: This is some input
Output: This is some output
`/* This is a comment */`

```
1  some code here;
2  x ← 0;
3  y ← 0;
4  if x > 5 then
5  |  x is greater than 5 ;                    // This is also a comment
6  else
7  |  x is less than or equal to 5;
8  end
9  foreach y in 0..5 do
10 |  y ← y + 1;
11 end
12 for y in 0..5 do
13 |  y ← y − 1;
14 end
15 while x > 5 do
16 |  x ← x − 1;
17 end
18 return Return something here;
```