CS-584 Midterm 1

10/23/24 Wyatt Blair

1)

1. Natural language processing is a subfield of AI which focuses on interpreting Natural language (i.e. a language spoken by humans) mathematically and algorithmically so it can be used in computer programs.
   a. Censoring: NLP can be used to identify malicious chat messages so troublesome users can be banned.
   b. Sentiment Analysis: determine whether users like or dislike a particular concept.
2. C: object detection
3. True

2) Tokens = $\{T[1:p_1-1], T[p_1+1:p_2-1], \ldots T[p_k+1:n]\}$

1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	"	C	o	m	p	u	t	a	t	i	o	n	a	l		l	i	n	g	u	i	s
1	t	i	c	s		i	s		t	h	e		i	n	t	e	r	s	e	c	t
2	i	o	n		o	f		a	r	t	i	f	i	c	i	a	l		i	n	t
3	e	l	l	i	g	e	n	c	e		a	n	d		h	u	m	a	n		l
4	a	n	g	u	a	g	e		s	t	u	d	y	"							

if we include the period

a. $p := \{14, 26, 29, 33, 46, 49, 60, 73, 77, 83, 92, 98\}$

b. tokens = $\{T[1:13], T[15:25], T[27:28], T[30:32], T[34:45], T[47:48], T[50:59], T[61:72], T[74:76], T[78:82], T[84:91], T[94:97], \}$

2.

a. Word tokenization is used in simpler NLP tasks and simply splits the strings into individual words using spaces / punctuation as delimeters.

b. sub-word tokenization splits each word into constituent parts. For example the word "catching" might be tokenized as: "catch" & "ing"

3) ☆☆ using this index to refer to a specific bigram so I don't need to rewrite them

1.

Bigrams = {"computational linguistics",
"linguistics is",
"is true",
"the intersection",
"intersection of",
"of artificial",
"artificial intelligence",
"intelligence and",
"and human",
"human language",
"language study"}

index
0
1
2
3
4
5
6
7
8
9
10

2.

index	0	1	2	3	4	5	6	7	8	9	10
count	1	1	1	1	1	1	1	1	1	1	1

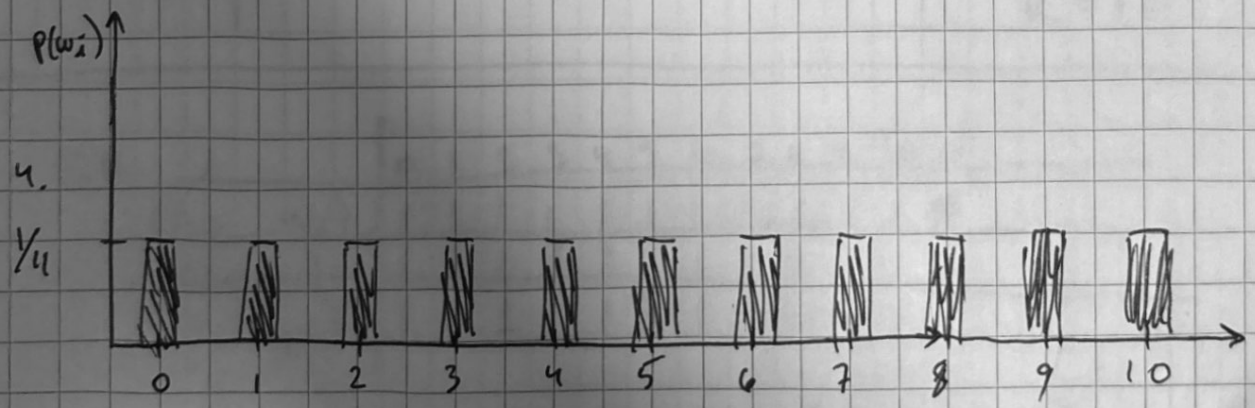
3.

i	0	1	2	3	4	5	6	7	8	9	10
$P(\omega_i)$	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11

$N = 11$

$P(\omega_i) = \frac{C(\omega_i)}{N}$

4.



4)

1.

a.

i	0	1	2	3	4	5	6	7	8	9	10
$C(w_{i-1}, w_i)$	1	1	1	1	1	1	1	1	1	1	1

b. Unigrams = {
"computational",
"linguistics",
"is",
"the",
"intersection",
"of",
"artificial",
"intelligence",
"and",
"human",
"language",
"study"
}

unigram index (j)

0	-1
1	0
2	1
3	2
4	3
5	4
6	5
7	6
8	7
9	8
10	9
11	10

3.

	-1	0	1	2	3	4	5	6	7	8	9	10
$j = i-1$	~~0~~	~~1~~	~~2~~	~~3~~	~~4~~	~~5~~	~~6~~	~~7~~	~~8~~	~~9~~	~~10~~	~~11~~
$C(w_j)$	1	1	1	1	1	1	1	1	1	1	1	1

number of unigrams
$K = 12$

c.

i	0	1	2	3	4	5	6	7	8	9	10	11
$C(w_{i-1}, w_i)$	1	1	1	1	1	1	1	1	1	1	1	1
$C(w_{i-1})$	1	1	1	1	1	1	1	1	1	1	1	1
$P(w_i \mid w_{i-1})$	1	1	1	1	1	1	1	1	1	1	1	

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

4)

2.

We want to find the t which maximizes $P(w_t|w_{t-1}) = \frac{C(w_{t-1}, w_t)}{C(w_{t-1})}$

where w_{t-1} = "artificial intelligence"

$$P(w_t|w_{t-1}) = \begin{cases} 0 \text{ if } t \neq 7 \\ 1 \text{ if } t = 7 \end{cases}$$ ← using bigram index

$\Rightarrow \max_t P(w_t|w_{t-1}) \rightarrow \boxed{t=7} \Rightarrow w_t = w_7 = \text{"and"}$

if we were meant to predict the next bigram and not the next unigram you could follow the same logic laid out above to determine

"and human" would be the next bigram