Quiz # 1
Natural Language Processing

Total Marks: 30 Course Code: CS-5840

Q1: Discuss shortly about the topic of your term paper that you have to submit at the end of this course. [5]
Solution: Depends on Student's project.

Q2: Calculate **precision** and **recall** for the following case study. Imagine there are 100 positive cases among 10,000 cases. You want to predict which ones are positive, and you pick 200 to have a better chance of catching many of the 100 positive cases. You record the IDs of your predictions, and when you get the actual results, you sum up how many times you were right or wrong. There are four ways of being right or wrong as follows:

TN / True Negative: case was negative and predicted **negative**
TP / True Positive: case was positive and predicted **positive**
FN / False Negative: case was positive but predicted **negative**
FP / False Positive: case was negative but predicted **positive**

After calculating the values for the four ways cited above, answer the following questions: [6]

1. What percent of your predictions were correct?
2. What percent of the positive cases did you catch?
3. What percent of positive predictions were correct?

Solution:

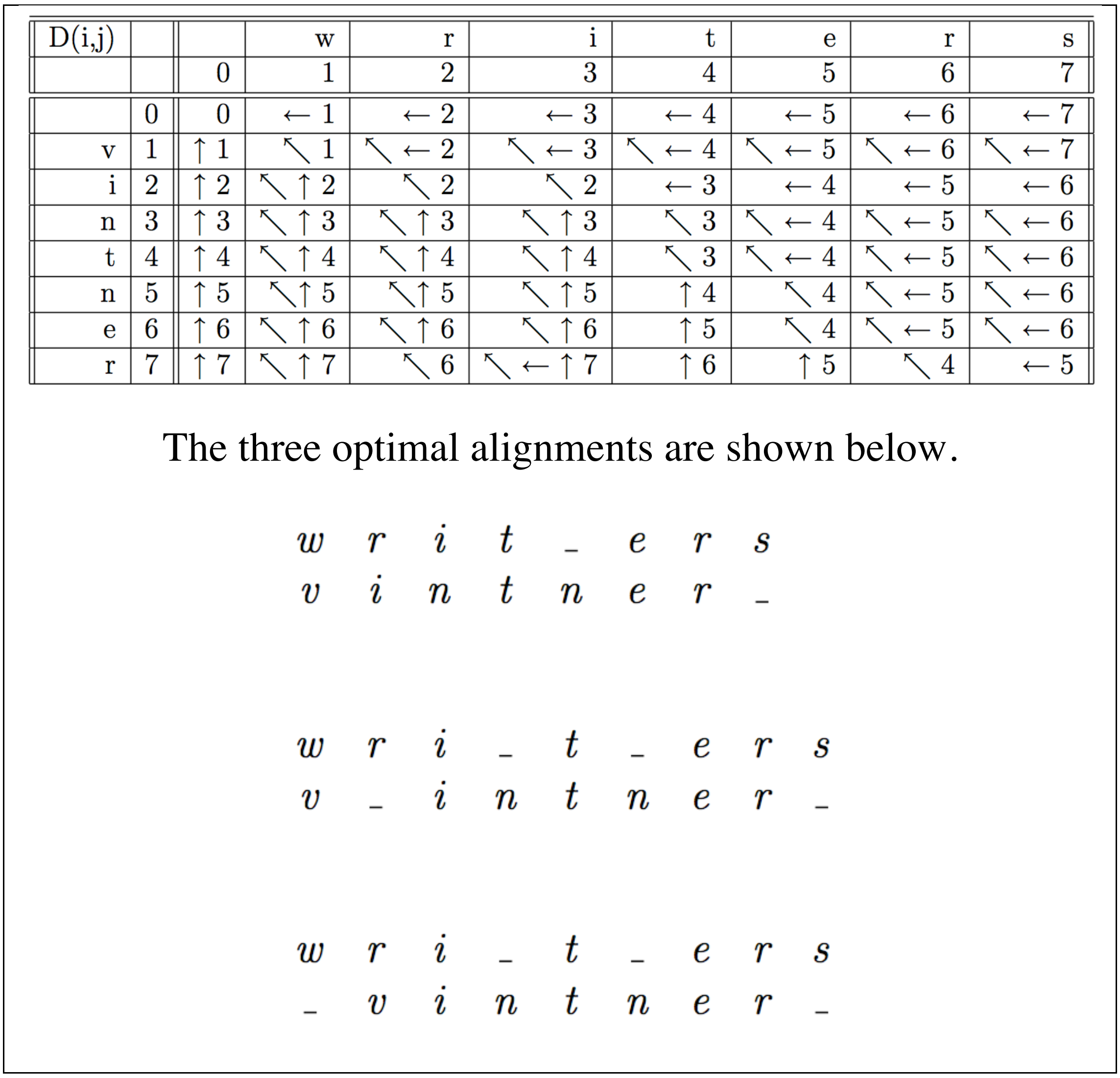
- Makes sense so far? Now you count how many of the 10,000 cases fall in each bucket, say:

	Predicted Negative	**Predicted Positive**
Negative Cases	TN: 9,760	FP: 140
Positive Cases	FN: 40	TP: 60

- Now, your boss asks you three questions:
- **What percent of your predictions were correct?**
  You answer: the "accuracy" was (9,760+60) out of 10,000 = 98.2%
- **What percent of the positive cases did you catch?**
  You answer: the "recall" was 60 out of 100 = 60%
- **What percent of positive predictions were correct?**
  You answer: the "precision" was 60 out of 200 = 30%

Q3: Apply the edit distance algorithm to modify the word "vintner" into "writers". Also apply the concept of back pointers through which you can provide the two given words in an alignment finally, along with the labels of operations (Insertion, deletion, substitutions). Assume the Levenshtein's proposal for the costs as 1,1,2 for the operations, respectively. [10]

Solution:



D(i,j)			w	r	i	t	e	r	s
		0	1	2	3	4	5	6	7
	0	0	← 1	← 2	← 3	← 4	← 5	← 6	← 7
v	1	↑ 1	↖ 1	↖ ← 2	↖ ← 3	↖ ← 4	↖ ← 5	↖ ← 6	↖ ← 7
i	2	↑ 2	↖ ↑ 2	↖ 2	↖ 2	← 3	← 4	← 5	← 6
n	3	↑ 3	↖ ↑ 3	↖ ↑ 3	↖ ↑ 3	↖ 3	↖ ← 4	↖ ← 5	↖ ← 6
t	4	↑ 4	↖ ↑ 4	↖ ↑ 4	↖ ↑ 4	↖ 3	↖ ← 4	↖ ← 5	↖ ← 6
n	5	↑ 5	↖↑ 5	↖↑ 5	↖ ↑ 5	↑ 4	↖ 4	↖ ← 5	↖ ← 6
e	6	↑ 6	↖ ↑ 6	↖ ↑ 6	↖ ↑ 6	↑ 5	↖ 4	↖ ← 5	↖ ← 6
r	7	↑ 7	↖ ↑ 7	↖ 6	↖ ← ↑ 7	↑ 6	↑ 5	↖ 4	← 5

The three optimal alignments are shown below.

```
w r i t _ e r s
v i n t n e r _
```

```
w r i _ t _ e r s
v _ i n t n e r _
```

```
w r i _ t _ e r s
_ v i n t n e r _
```

Q4: The following is the given corpus including <s>, </s> and punctuation marks as tokens.

<s> This is a list containing the tallest buildings in San Francisco : </s>
<s> The Transamerica Pyramid is the tallest building in San Francisco . </s>
<s> 555 California Street is the 2nd-tallest building in San Francisco . </s>

Calculate the detailed P_{KN} (probability using Kneser Ney smoothing) from the given corpus for the following cases. [9]

P_{KN}(Francisco|San) =?
P_{KN}(building|the tallest) =?
P_{KN}(building|is the 3rd-tallest) =?

P.T.O

Hint: Ney et al. [NEK94] estimate the discount value D based on the total number of n-grams occurring exactly once (n_1) and twice (n_2) [CG99] as $D = n_1/(n_1+2n_2)$. Stats of the given corpus are as below:

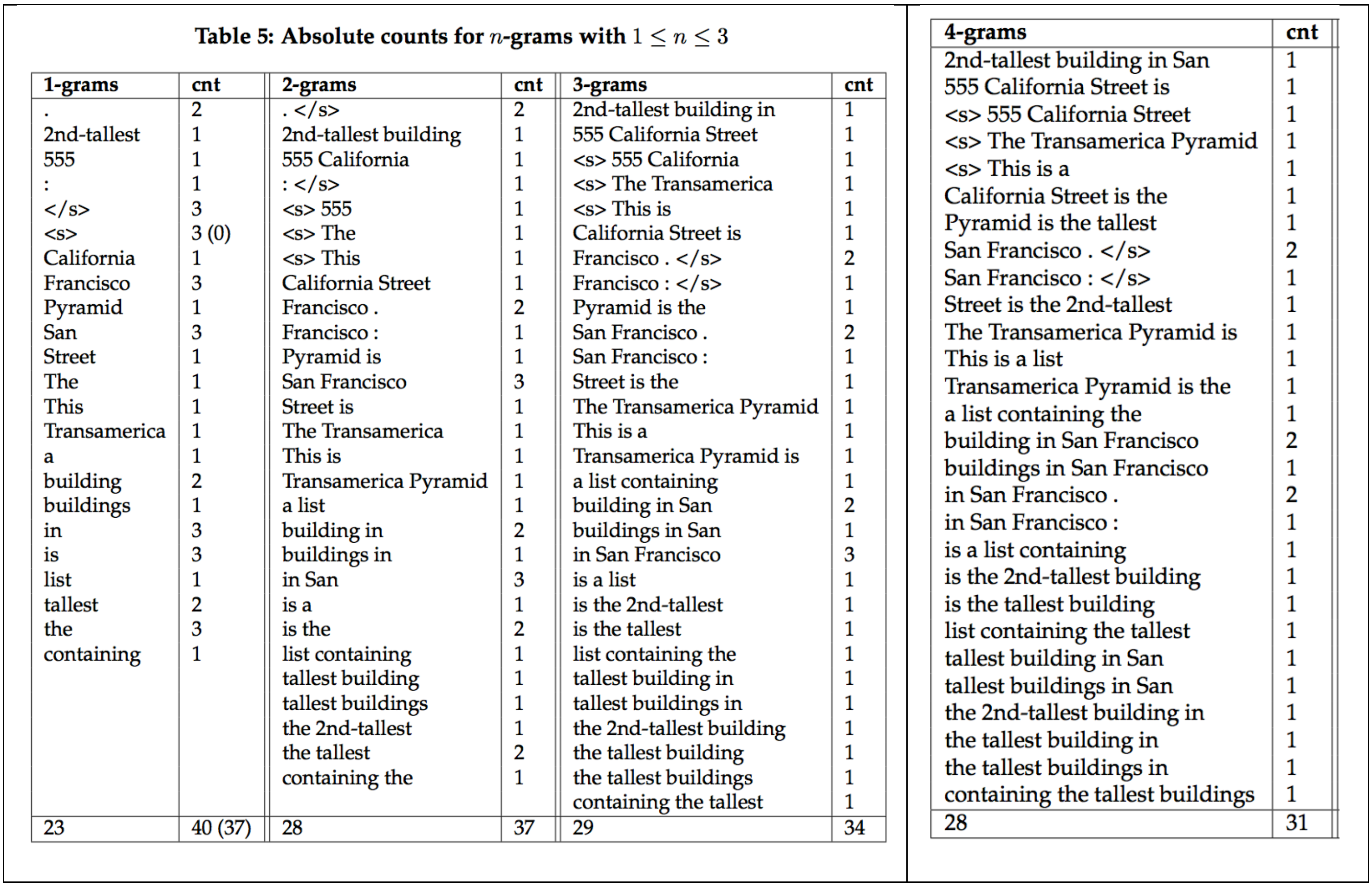


Table 5: Absolute counts for n-grams with $1 \leq n \leq 3$

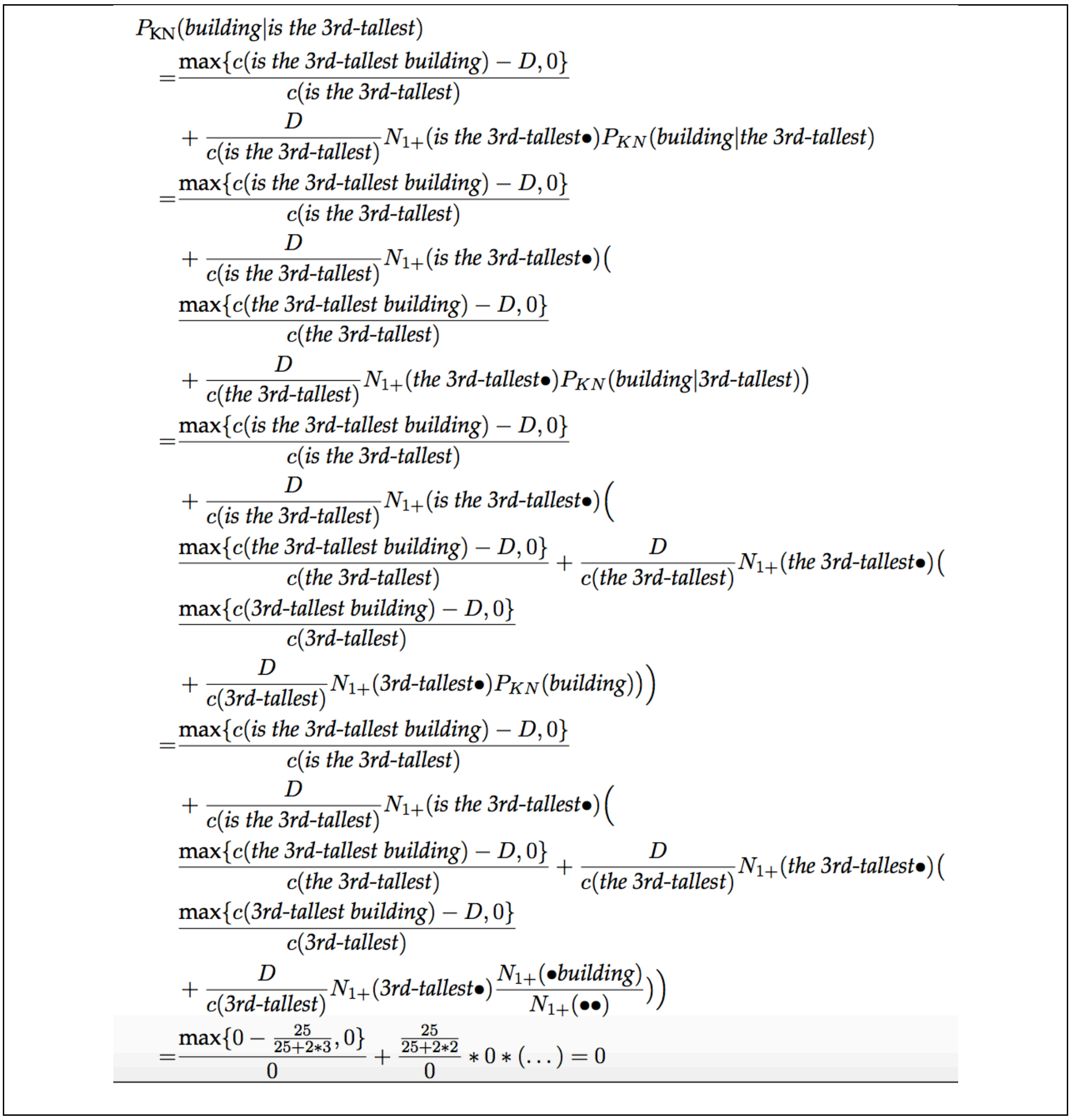
1-grams	cnt	2-grams	cnt	3-grams	cnt
.	2	. </s>	2	2nd-tallest building in	1
2nd-tallest	1	2nd-tallest building	1	555 California Street	1
555	1	555 California	1	<s> 555 California	1
:	1	: </s>	1	<s> The Transamerica	1
</s>	3	<s> 555	1	<s> This is	1
<s>	3 (0)	<s> The	1	California Street is	1
California	1	<s> This	1	Francisco . </s>	2
Francisco	3	California Street	1	Francisco : </s>	1
Pyramid	1	Francisco .	2	Pyramid is the	1
San	3	Francisco :	1	San Francisco .	2
Street	1	Pyramid is	1	San Francisco :	1
The	1	San Francisco	3	Street is the	1
This	1	Street is	1	The Transamerica Pyramid	1
Transamerica	1	The Transamerica	1	This is a	1
a	1	This is	1	Transamerica Pyramid is	1
building	2	Transamerica Pyramid	1	a list containing	1
buildings	1	a list	1	building in San	2
in	3	building in	2	buildings in San	1
is	3	buildings in	1	in San Francisco	3
list	1	in San	3	is a list	1
tallest	2	is a	1	is the 2nd-tallest	1
the	3	is the	2	is the tallest	1
containing	1	list containing	1	list containing the	1
		tallest building	1	tallest building in	1
		tallest buildings	1	tallest buildings in	1
		the 2nd-tallest	1	the 2nd-tallest building	1
		the tallest	2	the tallest building	1
		containing the	1	the tallest buildings	1
				containing the tallest	1
23	40 (37)	28	37	29	34

4-grams	cnt
2nd-tallest building in San	1
555 California Street is	1
<s> 555 California Street	1
<s> The Transamerica Pyramid	1
<s> This is a	1
California Street is the	1
Pyramid is the tallest	1
San Francisco . </s>	2
San Francisco : </s>	1
Street is the 2nd-tallest	1
The Transamerica Pyramid is	1
This is a list	1
Transamerica Pyramid is the	1
a list containing the	1
building in San Francisco	2
buildings in San Francisco	1
in San Francisco .	2
in San Francisco :	1
is a list containing	1
is the 2nd-tallest building	1
is the tallest building	1
list containing the tallest	1
tallest building in San	1
tallest buildings in San	1
the 2nd-tallest building in	1
the tallest building in	1
the tallest buildings in	1
containing the tallest buildings	1
28	31

Solution:

$$
\begin{aligned}
P_{\mathrm{KN}}(\mathit{Francisco}|\mathit{San}) =& \frac{\max\{c(\mathit{San\ Francisco}) - D, 0\}}{c(\mathit{San})} \\
&+ \frac{D}{c(\mathit{San})} N_{1+}(\mathit{San}\bullet) \frac{N_{1+}(\bullet \mathit{Francisco})}{N_{1+}(\bullet\bullet)} \\
=& \frac{\max\{3 - \frac{21}{21+2*5}, 0\}}{3} + \frac{\frac{21}{21+2*5}}{3} * 1 * \frac{1}{28} \\
\approx& \frac{2.32}{3} + \frac{0.68}{3} * 0.04 \approx 0.78
\end{aligned}
$$

$$
\begin{aligned}
&P_{\mathrm{KN}}(\mathit{building}|\mathit{the\ tallest}) \\
&= \frac{\max\{c(\mathit{the\ tallest\ building}) - D, 0\}}{c(\mathit{the\ tallest})} + \frac{D}{c(\mathit{the\ tallest})} N_{1+}(\mathit{the\ tallest}\bullet) P_{KN}(\mathit{building}|\mathit{the}) \\
&= \frac{\max\{c(\mathit{the\ tallest\ building}) - D, 0\}}{c(\mathit{the\ tallest})} \\
&\quad + \frac{D}{c(\mathit{the\ tallest})} N_{1+}(\mathit{the\ tallest}\bullet) \Big(\frac{\max\{c(\mathit{tallest\ building}) - D, 0\}}{c(\mathit{tallest})} \\
&\quad + \frac{D}{c(\mathit{tallest})} N_{1+}(\mathit{tallest}\bullet) P_{KN}(\mathit{building}) \Big) \\
&= \frac{\max\{c(\mathit{the\ tallest\ building}) - D, 0\}}{c(\mathit{the\ tallest})} \\
&\quad + \frac{D}{c(\mathit{the\ tallest})} N_{1+}(\mathit{the\ tallest}\bullet) \Big(\frac{\max\{c(\mathit{tallest\ building}) - D, 0\}}{c(\mathit{tallest})} \\
&\quad + \frac{D}{c(\mathit{tallest})} N_{1+}(\mathit{tallest}\bullet) \frac{N_{1+}(\bullet \mathit{building})}{N_{1+}(\bullet\bullet)} \Big) \\
&= \frac{\max\{1 - \frac{25}{25+2*3}, 0\}}{2} + \frac{\frac{25}{25+2*2}}{2} * 2 * \Big(\frac{\max\{1 - \frac{21}{21+2*5}, 0\}}{2} + \frac{\frac{21}{21+2*5}}{2} * 2 * \frac{2}{28} \Big) \\
&\approx \frac{0.19}{2} + \frac{0.81}{2} * 2 * \Big(\frac{0.32}{2} + \frac{0.68}{2} * 2 * 0.07 \Big) \approx 0.1 + 0.81 * 0.2 \approx 0.26
\end{aligned}
$$



$$
\begin{aligned}
&P_{\mathrm{KN}}(\textit{building}|\textit{is the 3rd-tallest})\\
&=\frac{\max\{c(\textit{is the 3rd-tallest building})-D,0\}}{c(\textit{is the 3rd-tallest})}\\
&\quad+\frac{D}{c(\textit{is the 3rd-tallest})}N_{1+}(\textit{is the 3rd-tallest}\bullet)P_{KN}(\textit{building}|\textit{the 3rd-tallest})\\
&=\frac{\max\{c(\textit{is the 3rd-tallest building})-D,0\}}{c(\textit{is the 3rd-tallest})}\\
&\quad+\frac{D}{c(\textit{is the 3rd-tallest})}N_{1+}(\textit{is the 3rd-tallest}\bullet)(\\
&\quad\frac{\max\{c(\textit{the 3rd-tallest building})-D,0\}}{c(\textit{the 3rd-tallest})}\\
&\quad+\frac{D}{c(\textit{the 3rd-tallest})}N_{1+}(\textit{the 3rd-tallest}\bullet)P_{KN}(\textit{building}|\textit{3rd-tallest}))\\
&=\frac{\max\{c(\textit{is the 3rd-tallest building})-D,0\}}{c(\textit{is the 3rd-tallest})}\\
&\quad+\frac{D}{c(\textit{is the 3rd-tallest})}N_{1+}(\textit{is the 3rd-tallest}\bullet)\Big(\\
&\quad\frac{\max\{c(\textit{the 3rd-tallest building})-D,0\}}{c(\textit{the 3rd-tallest})}+\frac{D}{c(\textit{the 3rd-tallest})}N_{1+}(\textit{the 3rd-tallest}\bullet)(\\
&\quad\frac{\max\{c(\textit{3rd-tallest building})-D,0\}}{c(\textit{3rd-tallest})}\\
&\quad+\frac{D}{c(\textit{3rd-tallest})}N_{1+}(\textit{3rd-tallest}\bullet)P_{KN}(\textit{building}))\Big)\\
&=\frac{\max\{c(\textit{is the 3rd-tallest building})-D,0\}}{c(\textit{is the 3rd-tallest})}\\
&\quad+\frac{D}{c(\textit{is the 3rd-tallest})}N_{1+}(\textit{is the 3rd-tallest}\bullet)\Big(\\
&\quad\frac{\max\{c(\textit{the 3rd-tallest building})-D,0\}}{c(\textit{the 3rd-tallest})}+\frac{D}{c(\textit{the 3rd-tallest})}N_{1+}(\textit{the 3rd-tallest}\bullet)(\\
&\quad\frac{\max\{c(\textit{3rd-tallest building})-D,0\}}{c(\textit{3rd-tallest})}\\
&\quad+\frac{D}{c(\textit{3rd-tallest})}N_{1+}(\textit{3rd-tallest}\bullet)\frac{N_{1+}(\bullet\textit{building})}{N_{1+}(\bullet\bullet)})\Big)\\
&=\frac{\max\{0-\frac{25}{25+2*3},0\}}{0}+\frac{\frac{25}{25+2*2}}{0}*0*(\dots)=0
\end{aligned}
$$