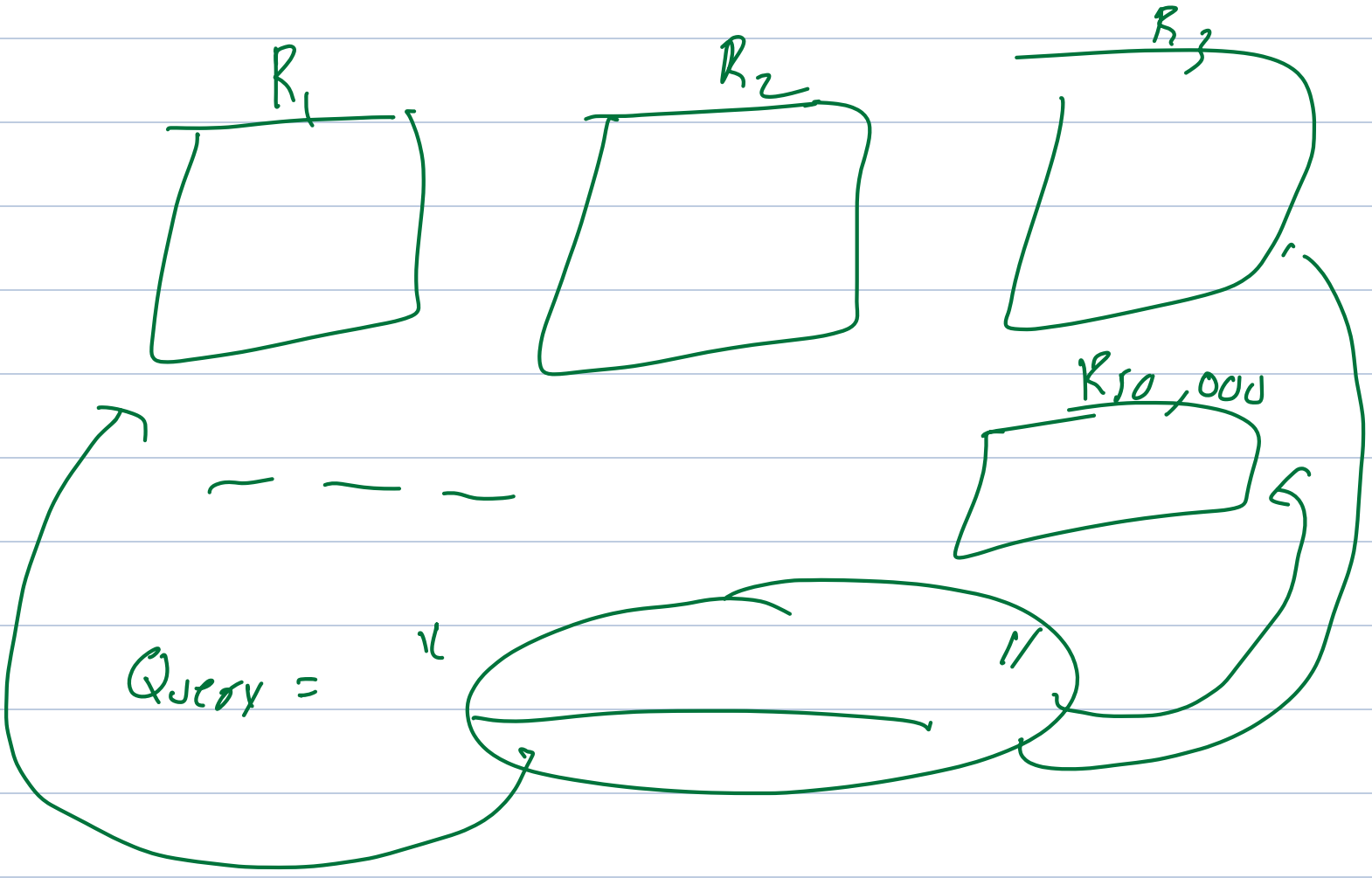
Language & topic modeling

Word2vec skipgram



Cosine (Query, R1)
(——, R2)
(——, R3)
(——, R4)

abc cde def
↓ ↓ ↓
[-,-,-,-] [-,-,-,-] [-,-,-,-]

R1 → 200 words

l = [0, 0, ---, 0]

l = l + model (each word)

l = l/200

centroid

Query →

→ origin of corona

R1 cos(o, c1) cos(of, c1) cos(corona, c1)

sum → 0.9

centroid 1

centroid 2

centroid 50, 0 v

R2 cos(o, c2) + cos(of, c2) + cos(corona, c2) ← 0.7

King = Man + Prince

Language & Topic

Generative

GenreqtiveAI

Generative world

Training data (corpus)

please find my CV | attached

→ P (attached | please find my CV)

Context | w2

P (w1 | context)

P (w2 | context)

p (Shakespeare | william)

Why drop

S1: the cat is small

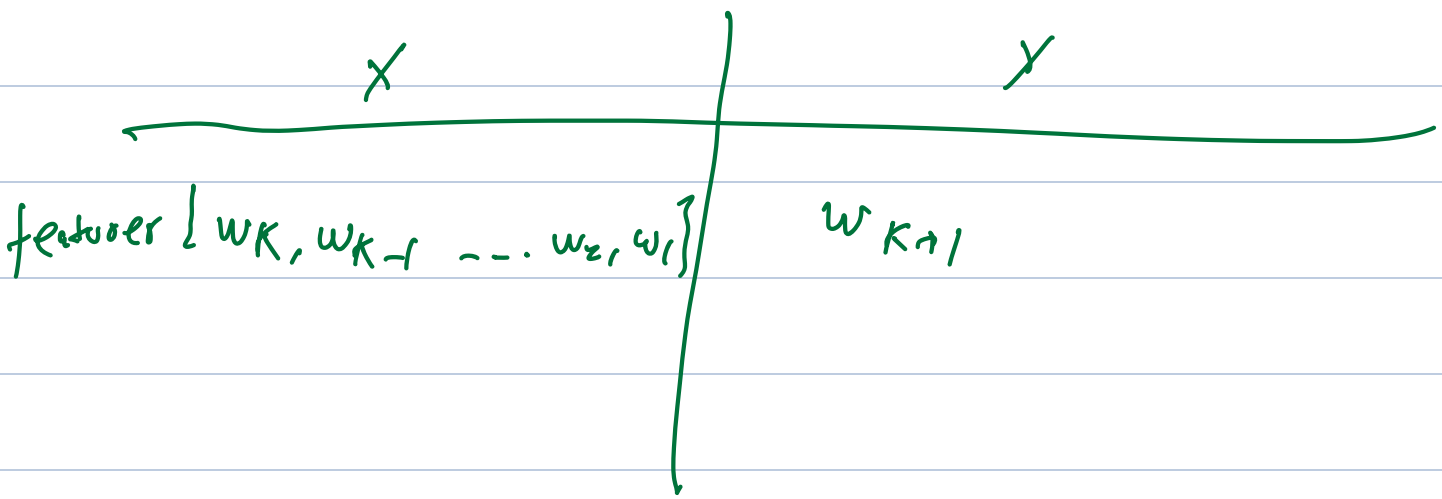
S2: small the cat is

order of word matter for

capital of france Paris

p(paris | capital of france

$p(w_{K+1} \mid w_K\ w_{K-1}\ \ldots\ w_2\ w_1)$



I

I hope $P(\text{hope} \mid I)$

I hope that $P(\text{that} \mid I \text{ hope})$

I hope that you $P(\text{you} \mid I \text{ hope that})$

$P(A, B, C, D) \rightarrow P(A) \cdot P(B|A) \cdot P(C|A,B) \cdot P(D|A,B,C)$

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

"its water is so transparent"

$P(\text{'} \quad \text{'}) = P(\text{its}) \cdot P(\text{water} \mid \text{its}) \cdot P(\text{is} \mid \text{its water}) \cdots$

$P(\text{transparent} \mid \text{its water is so})$

$$= \frac{\text{count}(\text{its water is so transparent})}{\text{count}(\text{its water is so})}$$

Markov Property

P(nextword | ———)

↳ K words

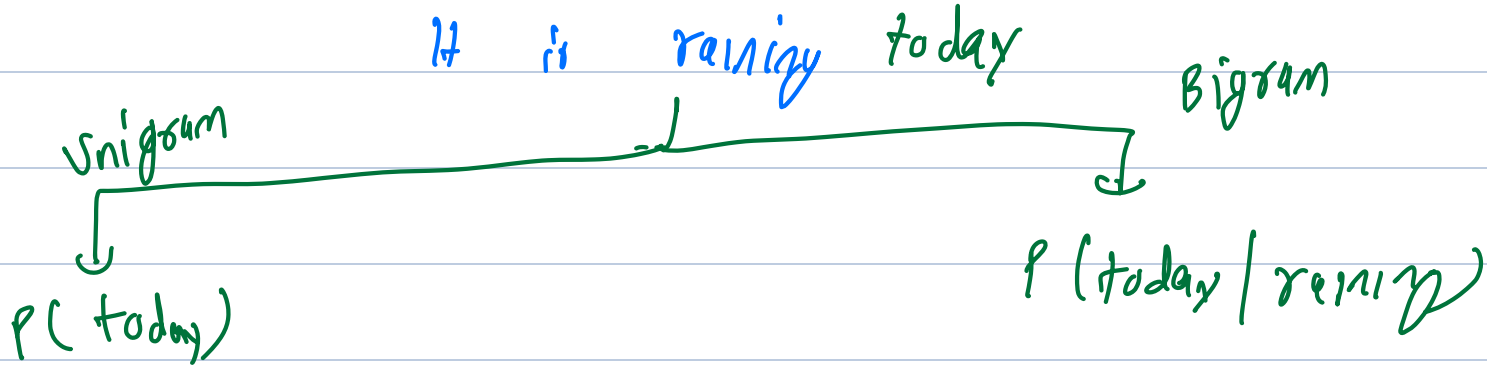
K=1 P(transparent | so)

K=2 P(transparent | is so)

K=1 ⇒ Unigram

K=2 ⇒ Bigram

K= n-1 ⇒ N-gram model



n= history

Unigram

P(I have a dream)

$= P(I) \cdot P(have) \cdot P(a)) \cdot P(dream)$

Bigram

$n=2$ <start>

P(I have a dream)

$P(I \mid \square) * P(have \mid I) * P(a \mid have) *$

$P(dream \mid a)$

<start>

Corpus

<start> I want a dream job </start>

<start> I have a dog </start>

<start> I have a dream company </start>

# of words

vocab size = 20

Unigram Table

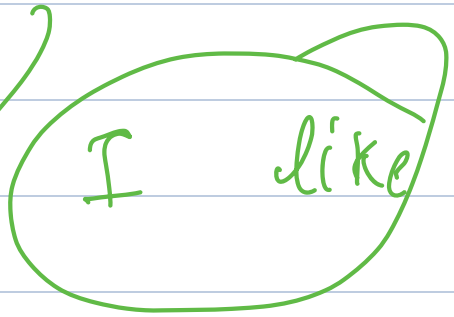
<start>	I	want	a	dream	job	have	dog	company
3/20	2/20	1/20	—	—	—	—		

$P(\text{I have a dream}) = P(\text{I}) \cdot P(\text{have}) \cdot P(\text{a}) \cdot P(\text{dream})$

$= \frac{}{} \cdot \frac{}{} \cdot \frac{}{} \cdot \frac{}{}$

Bigram Model

$P(w_2 \mid w_1) \Rightarrow P(\text{like} \mid \text{I})$



$P(\text{AI} \mid \text{gen})$

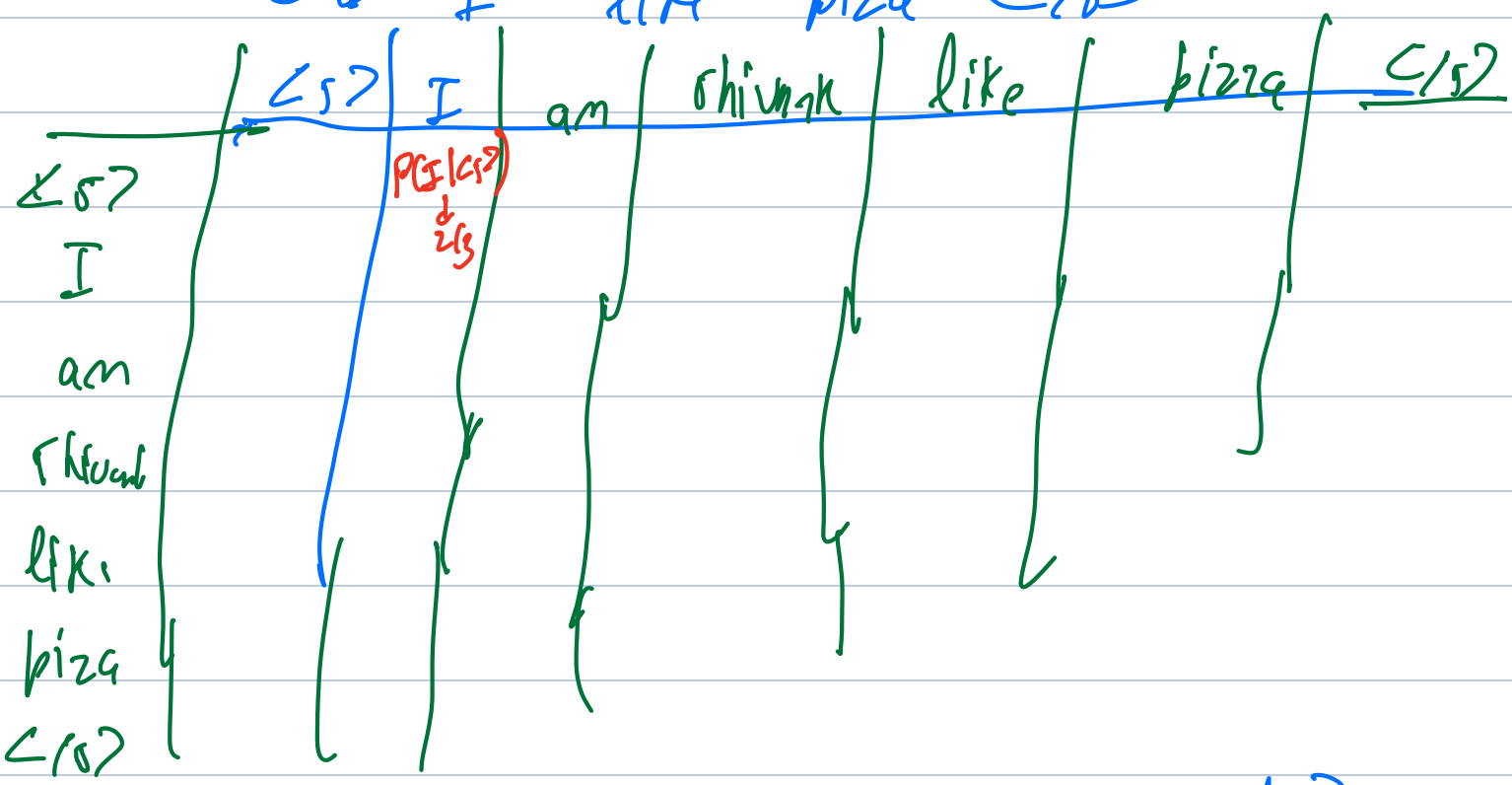
Generative AI

Corpus

<s> I am shivank </s>

<s> shivank I am </s>

<s> I like pizza </s>



Query ⇒ <s> shivank I like pizza</s>

P(shivank|<s>) · P(I|shivank) — — —

<s> I am shivank </s>

<s> shivank I am </s>

<s> I like pizza </s>