

Word2Vec

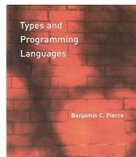
RecSys пайдалану жағдайы бойынша

- RS пайдалану жағдайы және енгізу-шығару арқылы оңай түсінуге болады:
 - Элемент кіріс ретінде, элемент шығыс ретінде: ұқсас(similar) және толықтырушы
 - (complementary)

Frequently bought together



+



+



Total price: **\$183.47**

Add all three to Cart

 Some of these items ship sooner than the others. [Show details](#)

- ☒ **This item:** Formal Semantics of Programming Languages by Glynn Winskel Paperback **\$70.00**
- ☒ Types and Programming Languages (The MIT Press) by Benjamin C. Pierce Hardcover **\$64.54**
- ☒ Practical Foundations for Programming Languages by Robert Harper Hardcover **\$48.93**

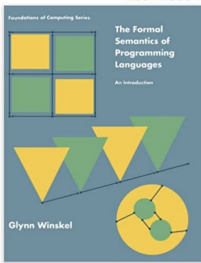
Input

Formal Semantics of Programming Languages

by Glynn Winskel (Author)

★★★★☆ 17 ratings

[Look inside](#)



Hardcover
\$45.07

Paperback
\$29.98 - \$70.00

Other Se
See all 5 version

☐ Buy used:
\$29.98

☒ Buy new:
\$70.00

In Stock.

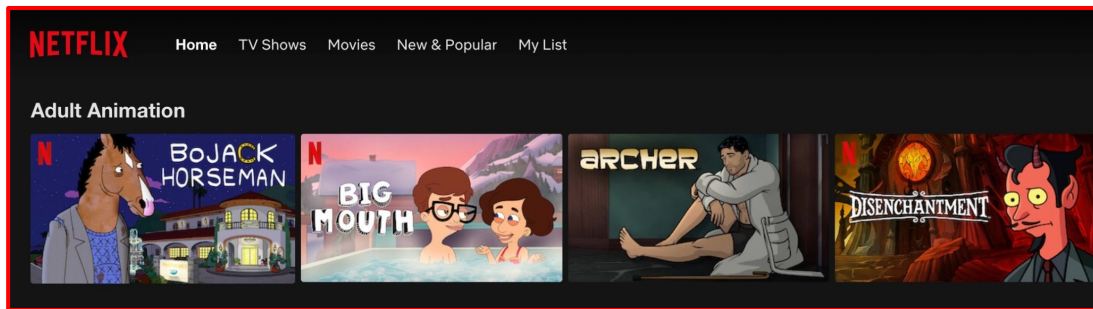
Ships from and sold by Amazon.com.

May be available at a lower price from other sellers, per Prime shipping.

RecSys пайдалану жағдайы бойынша

- RS пайдалану жағдайы және енгізу-шығару арқылы оңай түсінуге болады:
 - Элемент кіріс ретінде, элемент шығыс ретінде: ұқсас vs толықтырушы
 - Пайдаланушы - енгізу, элемент - шығару: «сіз үшін»


Input





RecSys пайдалану жағдайы бойынша


- RS пайдалану жағдайы және енгізу-шығару арқылы оңай түсінуге болады:
 - **Элемент - енгізу, элемент - шығару:** ұқсас vs толықтырушы
 - **Пайдаланушы - енгізу, элемент - шығару:** «сіз үшін»
 - **Пайдаланушы - енгізу, пайдаланушы - шығару:** «сіз білетін адамдар»
 - **Сеанс - енгізу, элемент - шығару:** «келесі не істейсіз?»


Input

1  Girls & Boys
Blur

▶  Rock 'n' Roll Star - Rem..
Oasis

3  Common People
Pulp

4  Lucky Man
The Verve

5  Linger
The Cranberries

6  Creep
E Radiohead

Pablo Honey

RecSys пайдалану жағдайы бойынша

- RS пайдалану жағдайы және енгізу-шығару арқылы оңай түсінуге болады:
 - **Элемент - енгізу, элемент - шығару:** ұқсас vs толықтырушы
 - **Пайдаланушы - енгізу, элемент - шығару:** «сіз үшін»
 - **Пайдаланушы - енгізу, пайдаланушы - шығару:** «сіз білетін адамдар»
 - **Сеанс - енгізу, элемент - шығару:** «келесі не істейсіз?»
 - **Элемент - енгізу, пайдаланушы - шығару:** «бұл затты кімге сатамыз?»

Input

*Жаңа фантастикалық
SaaS өнімі!*

*Негізгі интуиция:
егер сізге X ұнаса, сізге X -
қа ұқсас нәрселер де
ұнайды!*

«Ұқсас» нені білдіреді?

Ұқсастық және бейнелеу

Түйсік: ұқсастық - бұл «тиісті» кеңістіктегі «жақындық».

Фильмдерді екі өлшем бойынша
картаға түсірейік:

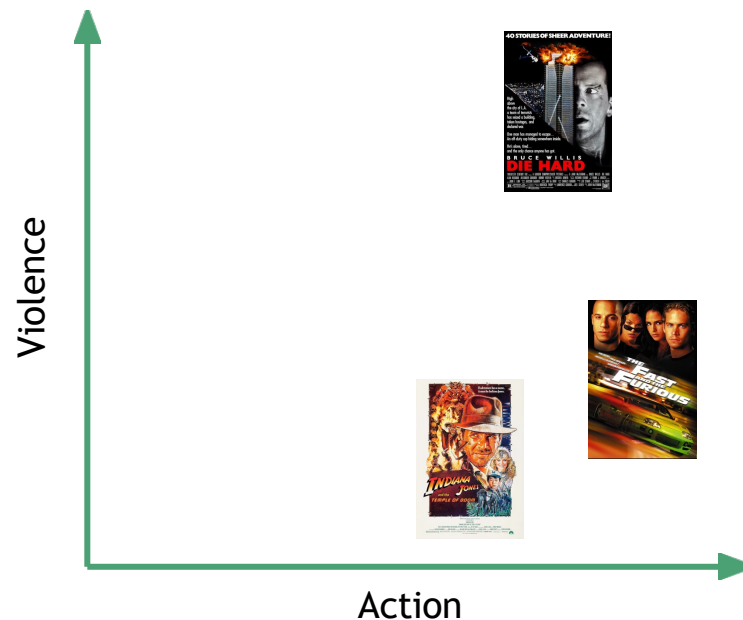
- Қанша әрекет бар?
- Қанша зорлық бар?

№1 бақылау: элементтер векторлар
ретінде

- Индиана Джонс: [3, 1]
- Жылдам және ашулы: [5, 2]
- Қиын өлу: [4, 4]

№2 бақылау: ұқсас фильмдер
кеңістікте жақын

- RecSys-ке оралу: егер сізге Индиана Джонс ұнаса, сізге FF, одан кейін Die Hard ұнайды.



Түйсік: ұқсастық - бұл «тиісті» кеңістіктегі «жақындық».

- Егер кеңістік негізгі ұғымдарды жақсы көрсетпесе, біз қиындыққа тап болдық!
- Машиналар векторларды түсінеді, бірақ барлық векторизациялар осы мағынада сәйкес кеңістікті анықтай бермейді.
- Мысалы, бір ыстық кодтауды қарастырайық:
 - «Мысық» «жыланға» қарағанда «итке» көбірек ұқсайды ма?
 - Ескерту: 1000 жануардан тұратын вектор қанша үлкен? Және қаншалықты сирек?

Dog

1	0	0
---	---	---

Snake

0	1	0
---	---	---

Cat

0	0	1
---	---	---

Түйсік: ұқсастық - бұл «тиісті» кеңістіктегі «жақындық».

- Егер кеңістік негізгі ұғымдарды жақсы көрсетсе, кеңістікте жақын элементтер ұқсас болады, бір-бірінен алыс орналасқан заттар соншалықты ұқсас емес.
- «Жақын» косинусты сипаттаудың көптеген әртүрлі тәсілдері бар
- қашықтық (немесе масштабталған векторлардағы нүктелік көбейтінді) ең көп таралған.
- Қорытынды: «ұқсастық туралы қорытынды» векторлық кеңістіктегі «жай» жақын көршіні іздеу

Cosine Similarity:

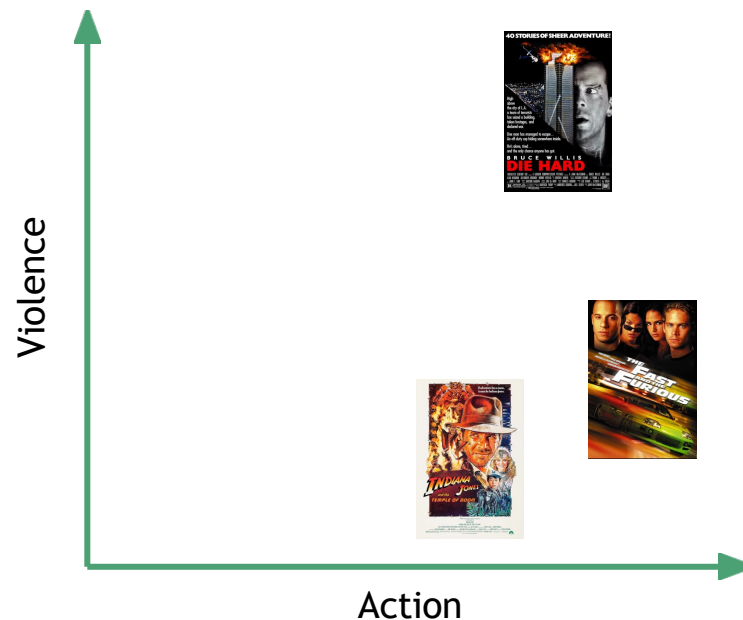
$$\text{cosine}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| |\mathbf{w}|} = \frac{\sum_{i=1}^N v_i w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}}$$

Dot Product

Түйсік: (негізгі) recSys GPS навигаторы сияқты

Фильмді ұсыну жүйелерін қарастырыңыз
(пайдаланушы элементінің жағдайы)

- 1-қадам: фильмдерді қолайлы кеңістікте көрсетіңіз

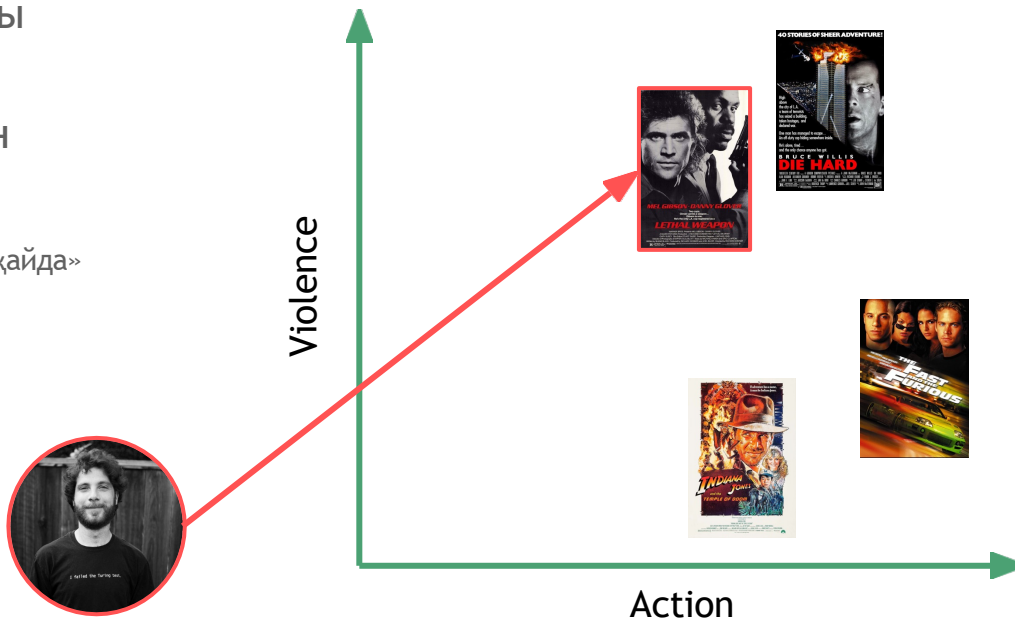


Түйсік: (негізгі) recSys GPS навигаторы сияқты

Фильмді ұсыну жүйелерін қарастырыңыз
(пайдаланушы элементінің жағдайы)

- 1-қадам: фильмдерді қолайлы кеңістікте көрсетіңіз
- 2-қадам: пайдаланушы орнын көрсетіңіз кеңістікте

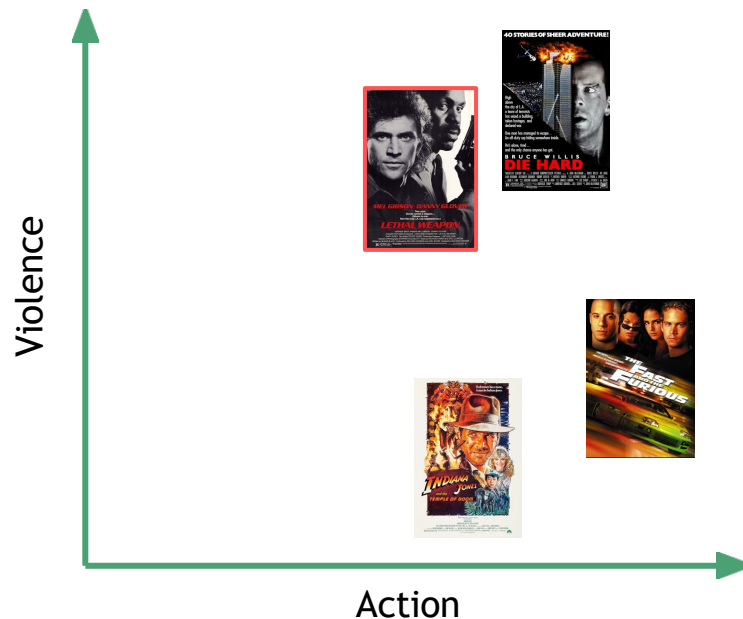
Мысалы, Джакопо «соңғы фильмі қайда»



Түйсік: (негізгі) recSys GPS навигаторы сияқты

Фильмді ұсыну жүйелерін қарастырыңыз (пайдаланушы элементінің жағдайы)

- 1-қадам: фильмдерді қолайлы кеңістікте көрсетіңіз
- 2-қадам: пайдаланушы орнын көрсетіңіз
- кеңістікте
- 3-қадам: пайдаланушыға ең жақын K элементін (KNN іздеу) ұсыныңыз!



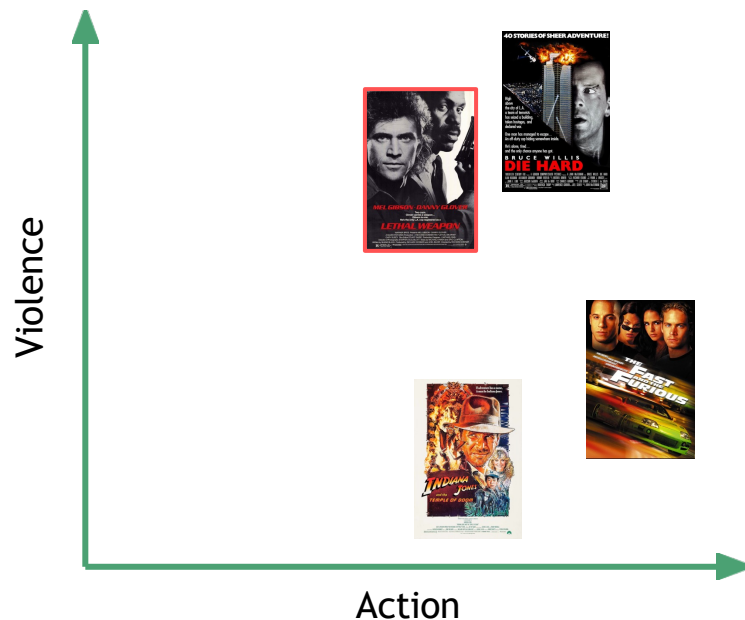
Түйсік: (негізгі) recSys GPS навигаторы сияқты

Нақты өмірлік мысалды қарастырайық:

Түйсік: пайдаланушылардың кеңістіктегі орнын білу арқылы (бұл жағдайда, Жер) біз олардың қалауы туралы көп айта аламыз!

recSys құрудағы табысымыздың үлкен бөлігі біздің кеңістіктегі өкілдіктің сапасына байланысты.

Пайдаланушылар мен элементтерді векторларға қалай салыстырамыз?



Өкілдіктер(representations) және деректер көздері

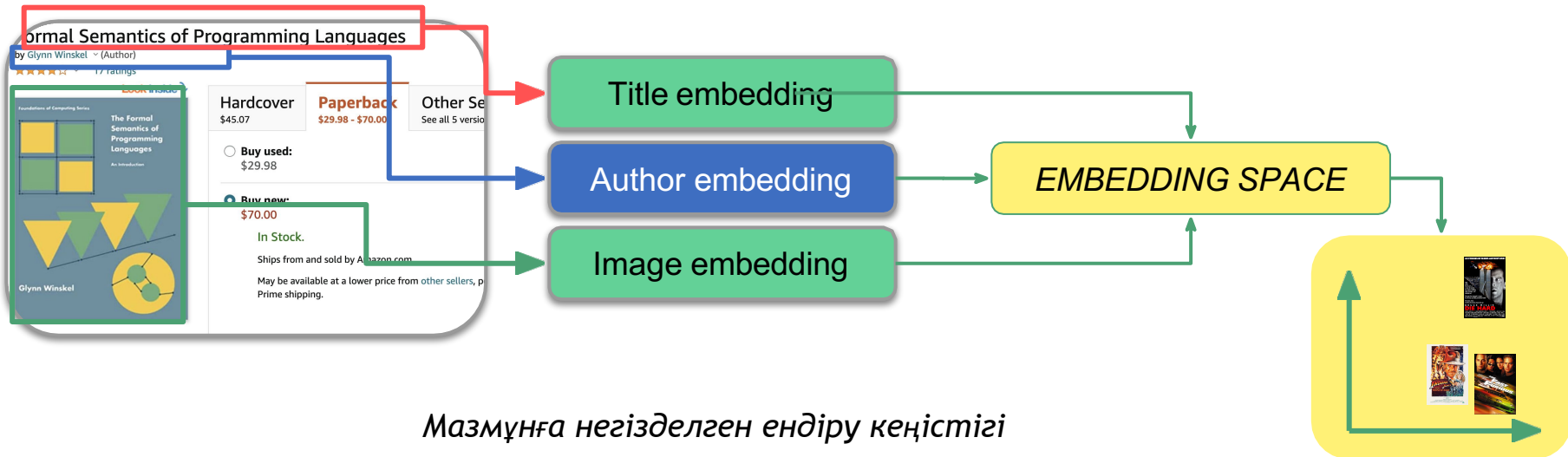
- **GOAL:** жақсы бейнелеу кеңістігін үйрену!
 - «Жақсы» кеңістік - бұл шын мәнінде ұқсас заттар жақын, ал бір-бірінен алыс заттар байланыссыз болатын кеңістік.
- Біз адамдардан Netflix-тегі барлық фильмдер үшін әрекетке қарсы зорлық-зомбылыққа және комедияға баға беруін сұрай алатын болсақ та, бұл мүмкін емес. :
 - Бір тонна қолмен жұмыс (мұны Amazon-тағы барлық кітаптар үшін жасауды елестетіп көріңіз!)
 - Қай жерде тоқтайтынымыз белгісіз: бізде актерлерге де өлшем болуы керек пе? Фильмнің ұзақтығы туралы не деуге болады? Өндіріс құны туралы не деуге болады? т.б.
- Біз әдетте оларды ажыратамыз **content-based (мазмұнға негізделген)** және **behavioral-based (мінез-құлыққа негізделген)**
 - Мазмұны: фильмнің тақырыбын, сценарийін, суреттерін, жанрын және т.б. талдау - яғни біздің каталогтағы бұл элемент туралы не білеміз?
 - Мінез-құлық: пайдаланушылардың элементтермен әрекетін талдау - егер пайдаланушы 1-ге А және В элементтерін ұнатса, содан кейін С-ны да ұнатса, С А және В-ге ұқсас деп есептей аламыз ба?

Өкілдіктер(representations) және деректер көздері

- **Мазмұнға негізделген көріністер** тек мақсатты нысандарымыздың «каталогын» қажет етеді: олардың сапасы суреттерді, мәтіндерді және санаттарды «жақсы» векторларға айналдыру мүмкіндігіне байланысты..
 - Кейде метадеректер жақсы емес (мысалы, фильмдер дұрыс емес санатталған!)
 - Кейде векторизация онша жақсы емес (мысалы, біздің тілдік модель аударма тілде жақсы жұмыс істемейді!)
- **Мінез-құлыққа негізделген өкілдіктер** «пайдаланушылардың» нақты деректерін талап етеді, е.g.:
 - Amazon-дан деректерді сатып алыңыз
 - Netflix-тен ағындық деректер
 - Spotify ойнату тізімінің деректері
 - т.б.
- **Note:** RecSys-тегі соңғы 20 жылдағы үлкен сабақ - мінез-құлыққа негізделген өкілдіктер жақсы бейнелерді жасауда таңқаларлық пайдалы (яғни адамдардың мінез-құлқында көптеген сигналдар бар!)

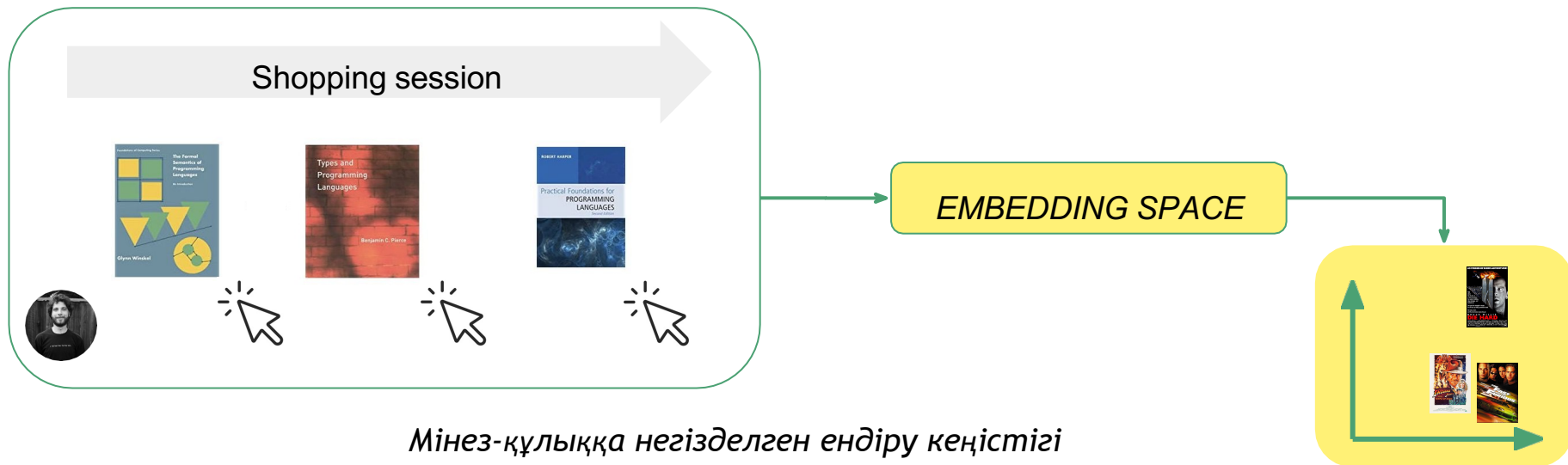
Өкілдіктер(representations) және деректер көздері

- Төмен өлшемді кеңістікке жоғары дәрежелі нысандарды (пайдаланушылар, элементтер, сөздер, т.б.) салыстыру процесін «енгізу»(embeddings) деп атаймыз:
 - Бір ыстық кодтау, сирек векторлар есіңізде ме? Енгізулер кішкентай (мыңдаған элементтер үшін 100 өлшемді) және тығыз!



Өкілдіктер(representations) және деректер көздері

- Төмен өлшемді кеңістікке жоғары дәрежелі нысандарды (пайдаланушылар, элементтер, сөздер, т.б.) салыстыру процесін «енгізу»(embeddings) деп атаймыз:
 - Бір ыстық кодтау, сирек векторлар есіңізде ме? Енгізулер кішкентай (мыңдаған элементтер үшін 100 өлшемді) және тығыз!



Word2Vec, Song2Vec, Everything2Vec

NLP аналогиясы: ұқсас заттар жиі бірге пайда болады

- **гипотеза: «ұқсас контексте кездесетін сөздердің мағыналары ұқсас»**
 - 1-мысал: егер екі кітап бір сауда сессиясында жиі қаралса, олар ұқсас болуы мүмкін!
 - 2-мысал: ойнату тізімдерінде екі ән жиі бірінен соң бірі болса, олар ұқсас болуы мүмкін!

Word Embeddings
Past, Present and Future

Сөздерді ендіруді үйренуге арналған рецепт («word2vec»)

- Жалпы аргументтің құрылымы :
 - Біз «жақсы» кеңістікті жасау үшін сөздерге арналған векторларды үйренуіміз керек
 - Ұқсас сөздер бір сөйлемде жиі кездеседі
 - Егер екі сөздің бірге пайда болуы мүмкін екенін көрсететін жіктеуіш үшін салмақ ретінде векторларды қолданатын болсақ, біз ұқсастықты кодтайтын, сондықтан жақсы кеңістікті тудыратын векторларды біле аламыз.!
- Басқаша айтқанда, дистрибуциялық гипотеза бізге ұқсастықтың прокси өлшемін береді: дистрибуциялық параметрлерде жақсы кірістірулер сөздердің ұқсастығы үшін жақсы көрініс болуы КЕРЕК.

Сөздерді ендіруді үйренуге арналған рецепт («word2vec»)

- Жалпы аргументтің құрылымы :
 - Біз «жақсы» кеңістікті жасау үшін сөздерге арналған векторларды үйренуіміз керек
 - Ұқсас сөздер бір сөйлемде жиі кездеседі
 - Егер екі сөздің бірге пайда болуы мүмкін екенін көрсететін жіктеуіш үшін салмақ ретінде векторларды қолданатын болсақ, біз ұқсастықты кодтайтын, сондықтан жақсы кеңістікті тудыратын векторларды біле аламыз.!
- Басқаша айтқанда, дистрибуциялық гипотеза бізге ұқсастықтың прокси өлшемін береді: дистрибуциялық параметрлерде жақсы кірістірулер сөздердің ұқсастығы үшін жақсы көрініс болуы КЕРЕК.

БОНУС ҰПТАЛЫ: бұл терең оқытудың ең сәтті қолданбаларында жұмыс істейді - сәйкес болжау тапсырмасын тапсаңыз, кез келген нәрсе үшін жақсы көріністерді біле аласыз!



Сөздерді ендіруді үйренуге арналған рецепт («word2vec»)

Corpus

CAT

IS

FURRY

Training the model weights

I

IS

FURRY

O

1

I

IS

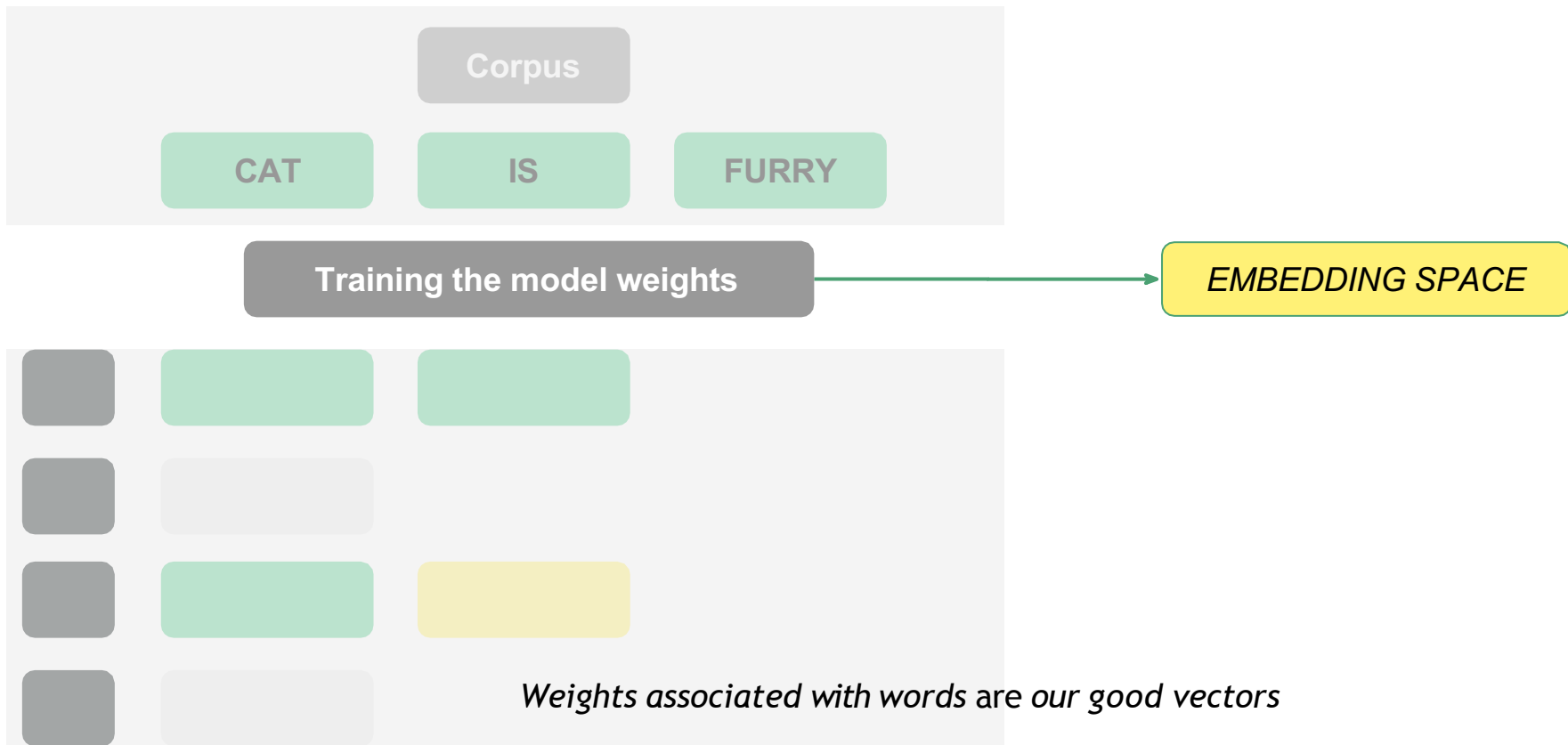
BERLIN

O

0

Turn sequential data into a prediction problem

A recipe for learning word embeddings (“word2vec”)



Болжау тапсырмасындағы сөз векторлары

- КОРПУС: «Түрлі мысық төсеніште»
- ТЕРЕЗЕ ҰЗЫНДЫҒЫ: 2
- МАҚСАТ: «мысық»
- КІРІСТІ ДАЙЫНДАУ, оң және теріс үлгілер

Target	Context	Label
cat	furry	1
cat	the	1
cat	is	1
cat	on	1

Target	Context	Label
cat	Berlin	0
cat	Jacopo	0
cat	ciao	0
cat	table	0

Болжау тапсырмасындағы сөз векторлары

- КОРПУС: «Түрлі мысық төсеніште»
- ТЕРЕЗЕ ҰЗЫНДЫҒЫ: 2
- МАҚСАТ: «мысық»
- КІРІСТІ ДАЙЫНДАУ, оң және теріс үлгілер

$$P_{\alpha}(w) = \frac{\text{count}(w)^{\alpha}}{\sum_{w'} \text{count}(w')^{\alpha}}$$

negative samples ($\alpha=0.75$)

Target	Context	Label
cat	furry	1
cat	the	1
cat	is	1
cat	on	1

Target	Context	Label
cat	Berlin	0
cat	Jacopo	0
cat	ciao	0
cat	table	0

Болжау тапсырмасындағы сөз векторлары

- Сөзді болжау мәселесін екілік жіктеу есебіне айналдырдық
Мәтінмәндік сөз мақсатты сөздің жанында пайда болуы мүмкін бе?
- Оқу мақсатымызды анықтайық:
Біз оң мысалдардан алынған (t, c) ұқсастығын барынша арттырғымыз келеді
Біз теріс мысалдардан алынған (t, c) ұқсастығын азайтқымыз келеді

$$L(\theta) = \sum_{(t,c) \in +} \log P(+|t, c) + \sum_{(t,c) \in -} \log P(-|t, c)$$

CHAPTER

6

Vector Semantics and Embeddings

荃者所以在鱼，得鱼而忘荃 Nets are for fish;
Once you get the fish, you can forget the net.
言者所以在意，得意而忘言 Words are for meaning;
Once you get the meaning, you can forget the words
庄子(Zhuangzi), Chapter 26

The aphorism that Lao Tzu is famous for refers mainly on its function. But

Болжау тапсырмасындағы сөз векторлары

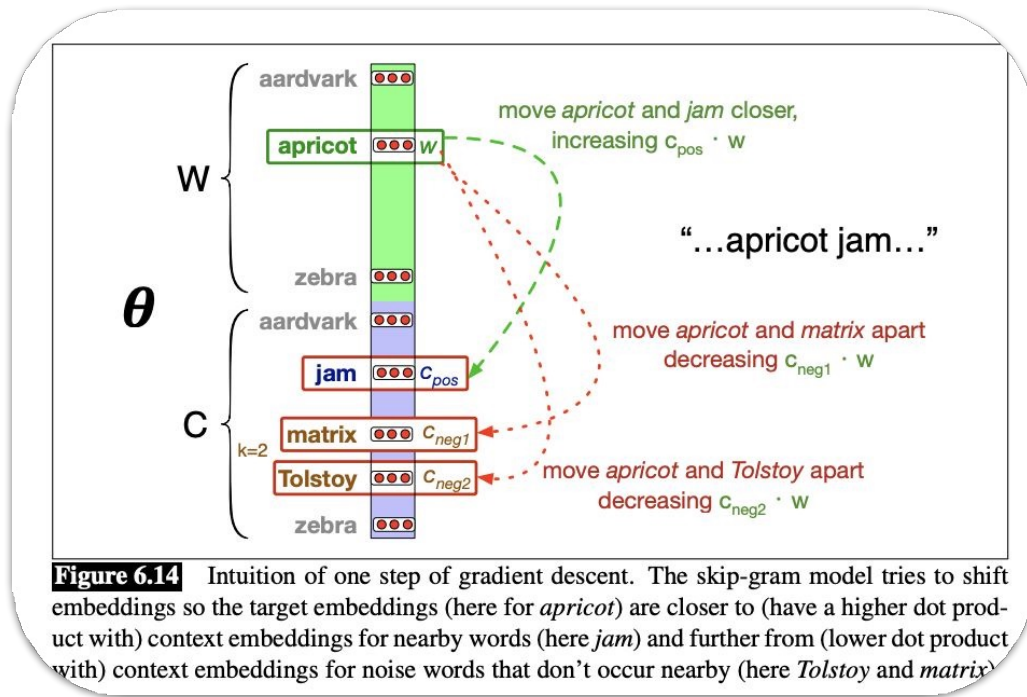
- Оқу мақсатымызды анықтайық :
 - Біз оң мысалдардан алынған (t, c) ұқсастығын барынша арттырғымыз келеді
 - Біз теріс мысалдардан алынған (t, c) ұқсастығын азайтқымыз келеді

The diagram illustrates the components of the log-likelihood function. A blue box labeled "Dot product" has an arrow pointing to the term $c \cdot t$ in the first equation. A green box labeled "Sigmoid" has an arrow pointing to the σ function in the same term. The equations are as follows:

$$= \log \sigma(c \cdot t) + \sum_{i=1}^k \log \sigma(-n_i \cdot t)$$
$$= \log \frac{1}{1 + e^{-c \cdot t}} + \sum_{i=1}^k \log \frac{1}{1 + e^{n_i \cdot t}}$$

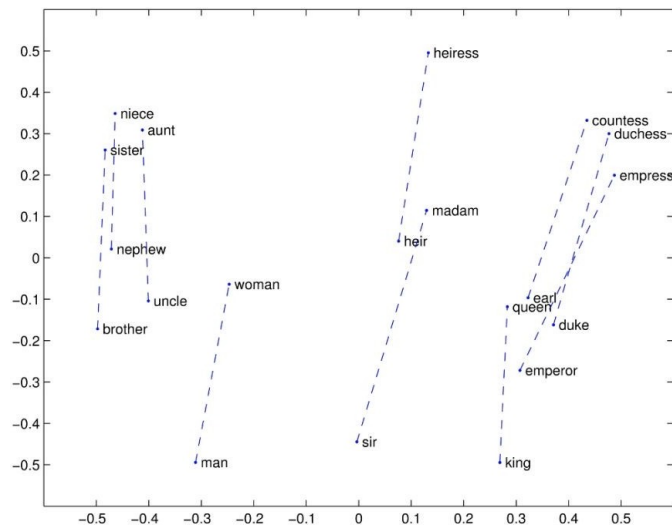
Болжау тапсырмасындағы сөз векторлары

- **Есіңізде болсын:** біз контекстік сөздермен сөздің нүктелік көбейтіндісін көбейтеміз, ал теріс таңдалған сөздермен сөздің нүктелік көбейтіндісін кішірейтеміз!
- **Жаттығу процедурасы :**
 - Сөздіктегі N сөзге арналған векторлардың кездейсоқ инициализациясы.
 - Әр қадам сайын қатысты сөздерді векторлық кеңістікке жақындатып, басқаларын одан әрі итеріңіз (градиенттің түсуін пайдалану арқылы).



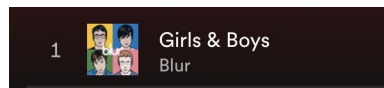
Бұл «жақсы» кеңістік пе?

- word2vec сөздер мен кейбір ұқсастық қатынастар арасындағы жақсы ұқсастықты алуға бейім - ешқандай адам белгілерінсіз / араласусыз!
- Жақсы дайындалған кірістіру кеңістігіне ие болғаннан кейін, векторлық ендірулер арасындағы ығысулар ұқсастықтарды шешу үшін пайдаланылуы мүмкін, мысалы: «ер: патша = әйелдер:?» (ханшайым)
- Бұл мүмкін, өйткені вектор('король') - вектор('ер') + вектор('әйел') векторына('король') жақын вектор болып табылады.

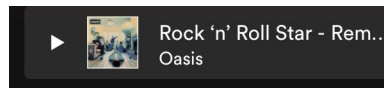


NLP-тен, RecSys-ке қайта оралу

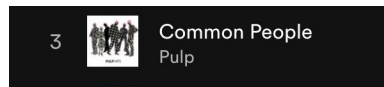
Есіңізде болсын: «сөйлемдегі сөздер» туралы бірдей интуицияны бізде мақсатты элементтердің (мысалы, ойнату тізімі, сауда сеанстары, т.б.) мәнді тізбектері болған кезде қолдануға болады.



CAT



IS



FURRY

Song2Vec



Book2Vec

Coding time!