Търсене и извличане на информация. Приложение на дълбоко машинно обучение

Стоян Михов

Лекция 5: Влагане на думи в многомерно векторно пространство.

План на лекцията

1. **Формалности за курса (5 мин)**
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (25 мин)
3. Смятане с вектори и матрици (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разредено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

Формалности

- Засега ще провеждаме занятията онлайн всяка сряда от 8:15 до 12:00 часа.
- Моля следете редовно обявите в Moodle за евентуални промени.
- Засега за лекциите ще използваме платформата Google meet: meet.google.com/hue-frfx-axb
- Днес ще използваме едновременно слайдове и бяла дъска. Моля следете съответния екран.
- Петата лекция се базира на глава 18 от първия учебник и глава 10 от втория учебник.

План на лекцията

1. Формалности за курса (5 мин)
2. **Вероятностно очакване, вариация, ковариация. Емпирично разпределение (25 мин)**
3. Смятане с вектори и матрици (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разредено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

Вероятностно очакване

- **Очакване на случайна величина** X означаваме с $\mathrm{E}[X]$ и дефинираме като $\mathrm{E}[X] = \sum_{x \in X(\Omega)} \Pr[X = x]\, x$

- Пример: Очакване на честен зар:
$$\sum_{i=1}^{6} \frac{1}{6} i = \frac{1+2+3+4+5+6}{6} = 3.5$$

- За някои случайни величини очакването може да е безкрайно: Петербургски пародокс: С вероятност $\frac{1}{2^n}$ стойността на величината е 2^n, за $n = 1,2,\ldots$

- **Очакване на функция на случайна величина** $f(X)$ означаваме с $\mathrm{E}[f(X)]$ и дефинираме като $\mathrm{E}[f(X)] = \sum_{x \in X(\Omega)} \Pr[X = x]\, f(x)$

- $H_X = -\,\mathrm{E}[\log(\Pr[X])]$

- Нека $X_1, X_2, \ldots, X_n$ са случайни величини над Ω. Тогава $\mathrm{E}[f(X_1, X_2, \ldots, X_n)] = \sum_{x_1 \in X_1(\Omega), x_2 \in X_2(\Omega), \ldots, x_n \in X_n(\Omega)} \Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n]\, f(x_1, x_2, \ldots, x_n)$

- Свойства:

  - $\mathrm{E}[af(X) + bg(Y)] = a\mathrm{E}[f(X)] + b\mathrm{E}[g(Y)]$

  - Ако X и Y са независими случайни величини, то $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$

Вариация

- **Вариацията (дисперсията) на случайна величина** X означаваме с $\mathrm{Var}[X]$ и дефинираме като $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$

- **Стандартна отклонение (девиация) на случайна величина** X означаваме с σ_X и дефинираме като $\sigma_X = \sqrt{\mathrm{Var}[X]}$

Свойства:

- $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$

- $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$

- Ако X и Y са независими то $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

Ковариация

- Ковариацията на две случайни величини X и Y означаваме с $\operatorname{Cov}(X, Y)$ и дефинираме като: $\operatorname{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$.

Свойства:

- $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$
- $\operatorname{Cov}(X + X', Y) = \operatorname{Cov}(X, Y) + \operatorname{Cov}(X', Y)$, $\operatorname{Cov}(aX, Y) = a\operatorname{Cov}(X, Y)$
- $\operatorname{Cov}(X, X) = \operatorname{Var}[X] \geq 0$
- $\operatorname{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$
- Ако X и Y са независими, то $\operatorname{Cov}(X, Y) = 0$

Емпирична функция на разпределение на вероятностна величина

- **Дефиниция**: Нека $X_1, X_2, \ldots, X_n$ са независими и идентично разпределени с X случайни величини. Нека сме наблюдавали (измерили) съответни стойности $x_1, x_2, \ldots, x_n$ за последователността от случайните величини $X_1, X_2, \ldots, X_n$. Емпиричното разпределение на случайните величини наричаме функцията на разпределение $\mathrm{Pr}_n[X = x] : x \mapsto \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i = x}$, където: $\delta_{X_i=x} = \begin{cases} 1 & \text{ако } X_i = x \\ 0 & \text{в противен случай} \end{cases}$.

- Емпирично очакване: $\mathrm{E}_n[X] = \sum_{x \in X(\Omega)} \mathrm{Pr}_n[X = x]\, x = \frac{1}{n}\sum_{i=1}^{n} x_i$

- $\mathrm{E}_n[f(X)] = \sum_{x \in X(\Omega)} \mathrm{Pr}_n[X = x]\, f(x) = \frac{1}{n} \sum_{x \in X(\Omega)} \sum_{i=1}^{n} \delta_{X_i = x}\, f(x) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$

- **Закони за големите числа**: (Няма да доказваме)
  - $\lim_{n\to\infty} \mathrm{Pr}_n[X = x] = \mathrm{Pr}[X = x]$ (Закон за големите числа на Борел);
  - $\mathrm{Pr}[\lim_{n\to\infty} \mathrm{E}_n[X] = \mathrm{E}[X]] = 1$ (Закон за големите числа на Колмогоров).

Емпирични оценки

Емпирична вариация, ковариация, ентропия и кросентропия на случайната величина:

- $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] \approx \frac{1}{n} \sum_{i=1}^{n} (x_i - \frac{1}{n} \sum_{j=1}^{n} x_j)^2$
- $\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] \approx \frac{1}{n} \sum_{i=1}^{n} (x_i - \frac{1}{n} \sum_{j=1}^{n} x_j)(y_i - \frac{1}{n} \sum_{j=1}^{n} y_j)$
- $H_X = -\,\mathrm{E}[\log \Pr_n[X]] = -\sum_{x \in X(\Omega)} \Pr_n[X = x]\ \log_2 \Pr_n[X = x] =$

  $= -\sum_{x \in X(\Omega)} \left(\frac{1}{n} \sum_{j=1}^{n} \delta_{X_j = x} \right) \log_2 \left(\frac{1}{n} \sum_{j=1}^{n} \delta_{X_j = x} \right) = -\frac{1}{n} \sum_{i=1}^{n} \log_2 \left(\frac{1}{n} \sum_{j=1}^{n} \delta_{X_j = x_i} \right)$
- $H_X(\Pr, \hat{\Pr}) = -\frac{1}{n} \sum_{i=1}^{n} \log_2 \hat{\Pr}[X = x_i]$

Пояснения за въпрос от миналата лекция — оценка на езиков модел

- Нека е даден езиков модел M с разпределение $\hat{\Pr}[x_n \mid x_1 x_2 \ldots x_{n-1}]$
- За да оценим действителното разпределение на езика, използваме достатъчно голям текст $x_1 x_2 \ldots x_m$ (често броят на думите в текста m е в порядък от милиони думи). Корпусът за оценяване не трябва да е използван за обучението на модела.
- **Перплексията** на езиковия модел M дефинираме като $2^{-\frac{1}{m}\sum_{n=1}^{m}\log_2 \hat{\Pr}[x_n \mid x_1 x_2 \ldots x_{n-1}]}$.
- $$-\frac{1}{m}\sum_{n=1}^{m}\log_2 \hat{\Pr}[x_n \mid x_1 x_2 \ldots x_{n-1}]$$ оценява крос-ентропията $H_X(\Pr \,||\, \hat{\Pr})$ между действителното разпределение на езика $\Pr$ и разпределението дадено от езиковия модел $\hat{\Pr}$.

- Ако езиковият модел е монограмен, то $-\frac{1}{m}\sum_{n=1}^{m}\log_2 \hat{\Pr}[x_n]$ е точно крос-ентропията на случайната величина X, еднакво разпределена с независимите случайни величини X_n, всяка от които ни дава съответно индекса на n-тата дума от текста x_n.

- Ако имаме немарковски езиков модел, разглеждаме едно наблюдение на случайната величина $X = x_1 x_2 \ldots x_m$. Тогава емпиричната крос-ентропия за едно наблюдение е:
$$H_X(\Pr\,||\,\hat{\Pr}) = -\log_2 \hat{\Pr}[x_1 x_2 \ldots x_{n-1}] = -\log_2 \prod_{n=1}^{m} \hat{\Pr}[x_n \,|\, x_1 x_2 \ldots x_{n-1}] = -\sum_{n=1}^{m}\log_2 \hat{\Pr}[x_n \,|\, x_1 x_2 \ldots x_{n-1}]$$
.
Това ни дава необходимия брой битове необходими за представяне на текста $x_1 x_2 \ldots x_m$ при използване на разпределението дадено от езиковия модел $\hat{\Pr}$ — **Entropy rate**.
Осреднено на дума са необходими $-\frac{1}{m}\sum_{n=1}^{m}\log_2 \hat{\Pr}[x_n \,|\, x_1 x_2 \ldots x_{n-1}]$ битове.

- За по-задълбочено изучаване на теория на информацията:
Курсът на Петър Митанкин “Основи на статистическата обработка на естествен език. Теория на информацията”. Ще се води зимния семестър на 2022 г.
Elements of Information Theory, Thomas M. Cover, Joy A. Thomas, John Wiley & Sons, 2012

Дефиниция на мярката за взаимна информация

- Нека X и Y са две случайни величини над вероятностно пространство Ω. Тогава мярката за взаимна информация $I(X;Y)$ на X и Y дефинираме:
$$I(X;Y) = D(\Pr[x,y]\,||\,\Pr[x]\ \Pr[y]) = \sum_{x\in X(\Omega),y\in Y(\Omega)} \Pr[x,y]\ \log_2 \frac{\Pr[x,y]}{\Pr[x]\Pr[y]}$$

- За удобство ще предполагаме, че $0 \log 0 = 0$ и $0 \log \dfrac{0}{0} = 0$.

- Когато случайните величини X и Y са независими, тяхното съвместно разпределение $\Pr[x,y]$ е равно на произведението на $\Pr[x]$ и $\Pr[y]$. Следователно взаимната информация е мярка за близостта на съвместното разпределение $\Pr[x,y]$ до неговата стойност, когато X и Y са независими, като близостта се измерва чрез релативната ентропията.

- По този начин $I(X;Y)$ може да се разглежда като мярка за количеството информация, която всяка една от величините може да предостави за другата.

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (25 мин)
3. **Смятане с вектори и матрици (20 мин)**
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разредено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

Означения на вектори и матрици

- Матриците ще бележим с главни удебелени букви — $\mathbf{W}, \mathbf{V}, \mathbf{A}, \mathbf{B}$, единична матрица — $\mathbf{I}$.
  $\mathbf{W}_{i,j}$ — елементът на ред i, стълб j в матрицата $\mathbf{W}$.
  $\mathbf{W}_{i,\bullet}$ — вектор ред i на матрицата $\mathbf{W}$.
  $\mathbf{W}_{\bullet,j}$ — вектор стълб j на матрицата $\mathbf{W}$.

- Векторите ще бележим с малки удебелени букви — $\mathbf{u}, \mathbf{v}, \mathbf{x}, \mathbf{y}$.
  $\mathbf{u}_i$ — i-тия елемент на $\mathbf{u}$.

- Ако не е указано друго, ще подразбираме вектор стълбове. Т.е. $\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u_2} \\ \vdots \\ \mathbf{u}_n \end{bmatrix}$.

- Векторите може да разглеждаме като матрици — $\mathbf{u} \in \mathbb{R}^{n \times 1}$.

- С $\mathbf{1}$ ще бележим вектор състоящ се само от единици.

Произведения на матрици и вектори

- Произведение на матрици: Нека $\mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{B} \in \mathbb{R}^{m \times k}$, тогава $\mathbf{C} = \mathbf{AB}$, ако $\mathbf{C} \in \mathbb{R}^{n \times k}$ и $\mathbf{C}_{i,j} = \sum_{l=1}^{m} \mathbf{A}_{i,l} \mathbf{B}_{l,j}$.

- Произведение на матрица с вектор:

  Нека $\mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{x} \in \mathbb{R}^{m \times 1}$, тогава $\mathbf{y} = \mathbf{Ax}$, ако $\mathbf{y} \in \mathbb{R}^{n \times 1}$ и $\mathbf{y}_i = \sum_{l=1}^{m} \mathbf{A}_{i,l} \mathbf{x}_l$.

  Нека $\mathbf{u} \in \mathbb{R}^{n \times 1}$, тогава вектора ред $\mathbf{v} = \mathbf{u}^\top \mathbf{A}$, ако $\mathbf{v} \in \mathbb{R}^{1 \times m}$ и $\mathbf{v}_i = \sum_{l=1}^{n} \mathbf{u}_l \mathbf{A}_{l,i}$.

- Скаларно произведение на два вектора: Нека $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{n \times 1}$ тогава $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^\top \mathbf{v} = \sum_{l=1}^{n} \mathbf{u}_l \mathbf{v}_l$.

- Диадно (тензорно, външно) произведение на два вектора: Нека $\mathbf{x} \in \mathbb{R}^{n \times 1}, \mathbf{y} \in \mathbb{R}^{m \times 1}$, тогава $\mathbf{x} \otimes \mathbf{y} = \mathbf{x}\mathbf{y}^\top = \mathbf{C}$, ако $\mathbf{C} \in \mathbb{R}^{n \times m}$ и $\mathbf{C}_{i,j} = \mathbf{x}_i \mathbf{y}_j$.

Градиент и Якобиан

- Нека $f : \mathbb{R}^n \to \mathbb{R}$. **Градиент** на f наричаме вектора:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \frac{\partial}{\partial x_2} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}) \end{bmatrix}.$$

- Нека $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ **Якобиан** на $\mathbf{f}$ наричаме матрицата:

$$\nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_m(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \nabla_{\mathbf{x}} f_1(\mathbf{x})^\top \\ \vdots \\ \nabla_{\mathbf{x}} f_m(\mathbf{x})^\top \end{bmatrix}.$$

Свойства

- Нека $\mathbf{x} \in \mathbb{R}^{n\times 1}$. Тогава $\dfrac{\partial}{\partial \mathbf{x}}\mathbf{x} = \mathbf{I}$.
- Нека $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{x} \in \mathbb{R}^{n\times 1}$. Тогава $\dfrac{\partial}{\partial \mathbf{x}}\mathbf{A}\mathbf{x} = \mathbf{A}$.
- Нека $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{n\times 1}$. Тогава $\dfrac{\partial}{\partial \mathbf{u}}\mathbf{u}^\top\mathbf{v} = \mathbf{v}$ и $\dfrac{\partial}{\partial \mathbf{v}}\mathbf{u}^\top\mathbf{v} = \mathbf{u}$.
- Нека $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{x} \in \mathbb{R}^{n\times 1}$. Тогава $\dfrac{\partial}{\partial \mathbf{x}}\mathbf{x}^\top\mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

Ковариационна матрица

- Ковариационна матрица на вектор от случайни величини $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ е матрица $\mathbb{R}^{n \times n}$, която означаваме с $\mathbf{C}[\mathbf{X}]$ и дефинираме като: $\mathbf{C}[\mathbf{X}] = \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^\top]$, т.е. $\mathbf{C}[\mathbf{X}]_{i,j} = \mathrm{Cov}(X_i, X_j)$

Свойство:

- $\mathbf{C}[\mathbf{X}] = \mathrm{E}[\mathbf{X}\mathbf{X}^\top] - \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{X}]^\top$

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (25 мин)
3. Смятане с вектори и матрици (20 мин)
4. **Семантично разширяване на заявката (10 мин)**
5. Влагане на думи във многомерно разредено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

Недостатъци при търсене базирано на съвпадение на ключови думи

- На упражнението видяхме пример за нерелевантно ранкиране при използване на $\mathrm{tf} \bullet \mathrm{idf}$ тегла (*“Румъния вирус”*).

- Задачата е да се удовлетвори информационната потребност, а не да се броят съвпадения на срещания на ключови думи между заявката и документа.

Пример:
Заявка: *“добра застраховка за кола”*
Релевантен документ: *“идеалното автомобилно каско”*,
Нерелевантен документ :*“добра застраховка живот покрива инциденти с кола”*

- Търсене базирано само на съвпадение на ключови думи в много случай връща нерелевантни и изпуска релевантни резултати.

- Следващата цел е да се реализира търсене по смисъл — т.е. по семантична близост.

Семантично разширяване на заявката

- Използване на семантичен речник.

    - Експертно съставен семантичен речник
      Пример: WordNet съдържа синонимни, антонимни, меронимни и хипонимни и други семантични връзки между думите.

    - Автоматично съставен семантичен речник — на базата на съвместно срещане на думите в документите:
      Пример: *Отидохме да берем ябълки и круши.* ===> термовете *ябълки* и *круши* са близки.

- Ключовите думи от заявката се разширяват със семантично свързани термове — също като при толерантното търсене.

Проблеми при семантично разширяване на заявката

- Експертно съставените речници са непълни, трудно се поддържат и не обхващат новите термини.

- Автоматично съставените семантични речници съдържат много шум (нерелевантност и неточност).

- Много от релациите са валидни само в определени контексти, което води до нерелевантно разширяване; (*маса ≈ тегло, маса от дърво ≉ тегло от дърво*).

- Ще разгледаме алтернативно решение чрез влагане на думите в “семантично” векторно пространство.

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (25 мин)
3. Смятане с вектори и матрици (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. **Влагане на думи във многомерно разредено векторно пространство (10 мин)**
6. Влагане на термовете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

Документно представяне във многомерно векторно пространство с “one hot” вектори

- В мултиномния документен модел на всеки документ съпоставяме $|V|$-мерен вектор d, в който на позиция t записваме броя на срещанията на съответния терм.

- Дефинираме за всеки терм с индекс t съответен **“one hot”** $|V|$-мерен вектор, $\mathbf{w}^{t_k} \in \{0,1\}^{|V|}$, който се състои от $|V| - 1$ нули и една единица на позиция k. Т.е.
$$\mathbf{w}_i^{t_k} = \begin{cases} 1 & \text{ако } i = k \\ 0 & \text{в противен случай} \end{cases}$$

$$\mathbf{w}^{t_k} = \begin{pmatrix} 0, & 0, & \ldots, & 0, & \underset{\substack{\uparrow \\ k}}{1,} & 0, & \ldots, & 0 \end{pmatrix}^{\top}$$

- В такъв случай получаваме: $d = \sum_{j=1}^{L_d} \mathbf{w}^{t_j}$

- Други документни представяния също могат да се разглеждат като получени от влагане на думи в многомерно векторно пространство. Например при някои варианти на $\mathrm{tf} \bullet \mathrm{idf}$ теглата.

- Във тези случаи векторите, които съпоставяме на термовете съдържат ненулева стойност единствено на позицията, която съотвества на терма.

- Документното представяне получаваме като сумираме (или акумулираме по друг начин) векторите, съответстващи на термовете от документа.

- Векторите, съответстващи на различни думи са ортогонални.

- Векторите, съответстващи на документите са силно разредени — размерността им е $|V|$, често над 100000, а броя на ненулевите стойности е по-малък от L_d — около 1000.

- <u>**Проблем**</u>: Няма никаква връзка между семантичната близост между термовете и техните векторни представяния. Следователно, векторното представяне на документа изцяло зависи от това, какви точно термове са използвани.

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (25 мин)
3. Смятане с вектори и матрици (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разредено векторно пространство (10 мин)
6. **Влагане на термовете в контекстно пространство (15 мин)**
7. Ранкиране на документи в контекстно пространство (5 мин)

Алтернативен подход за векторно представяне на думи и документи

- **Дистрибутивна семантика**: Значението на дадена дума се определя от думите, които често се срещат около нея.
  - *“You shall know a word by the company it keeps”* (Firth 1957)
- В тълковните речници значението се определя с примери за използването на думата.
- Пример: ***Зад храста се показа малък косматпирентил с вирната опашка***.

- Контекстът на дадена дума са думите, които са около нея — в рамките на параграф, изречение или фиксиран по размер прозорец.
- Две думи ще считаме за семантично свързани, ако често се срещат в един и същ контекст.
- Чрез статистически анализ върху контекстите на срещанията на думите определяме тяхната семантична близост.

Пример: Матрица на срещанията на терм в контекст

К1: Иван кара кола. Иван купи кола.
К2: Мария купи колело. Мария кара колело.
К3: Иван обича кола. Мария обича колело.
К4: Иван обича Мария.

Брой срещания на терма в съответния контекст

	К1	К2	К3	К4
Иван	2	0	1	1
Мария	0	2	1	1
кара	1	1	0	0
купи	1	1	0	0
обича	0	0	2	1
кола	2	0	1	0
колело	0	2	1	0

Пример: Матрица на съвместните срещания

Иван кара кола . Иван купи кола .
Мария купи колело . Мария кара колело .
Иван обича кола . Мария обича колело .
Иван обича Мария .

Брой срещания на терма в прозорец около съответната дума

	Иван	Мария	кара	купи	обича	кола	колело
Иван	0	0	1	1	2	0	0
Мария	0	0	1	1	2	0	0
кара	1	1	0	0	0	1	1
купи	1	1	0	0	0	1	1
обича	2	2	0	0	0	1	1
кола	0	0	1	1	1	0	0
колело	0	0	1	1	1	0	0

Терм / контекст матрица

- Нека V е наредено множество от термове и C е наредено множество от контексти. Нека функцията $f : V \times C \to \mathbb{R}$ е мярка за свързването на даден терм с даден контекст. Тогава дефинираме матрицата терм / контекст $M^f \in \mathbb{R}^{|V|\times|C|}$ като $M^f_{i,j} = f(t_i, c_j)$, където $t_i \in V$ е i-тия терм в V и $c_j \in C$ е j-тия контекст в C.

- На терм t_i съпоставяме съответния вектор ред на матрица: $t_i \mapsto M^f_{i,\bullet}$

- Близост или подобие между термовете t_i, t_k дефинираме:
$$\mathrm{sim}_{\cos}(t_i, t_k) = \cos(M^f_{i,\bullet}, M^f_{k,\bullet}) = \frac{M^f_{i,\bullet} \cdot M^f_{k,\bullet}}{|M^f_{i,\bullet}||M^f_{k,\bullet}|}$$

- Възможни са и други мярки за подобие но косинусовата близост е най-често и най-успешно използваната.

Мярка за свързването на терм с контекст

- Най-простата мярка е броя на срещанията:
  $f(t_i, c_j) = \#(t_i, c_j)$, където с $\#(t_i, c_j)$ означаваме броя на срещанията на терма t_i в контекста c_j.

- Често се използва честотата на срещанията — броя нормализиран към сумата от всички срещания:
  $f(t_i, c_j) = \dfrac{\#(t_i, c_j)}{|D|}$, където с $|D| = \sum_{t \in V, c \in C} \#(t, c)$.
  В такъв случай имаме, че $f(t_i, c_j) = \Pr_{MLE}[t_i, c_j]$.

- Недостатък на броя на срещанията е, че се получават много високи стойности за често срещани термове като предлози, определителни думи и други.

- Най-добри резултати се получават с използване на поточкова мярка за взаимна информация:
  $$\mathrm{PMI}(t; c) = \log \frac{\Pr[t, c]}{\Pr[t]\ \Pr[c]} = \log \frac{\#(t, c)\,|D|}{\#(t, \bullet)\,\#(\bullet, c)}.$$
  (Предполагаме, че ако $\#(t, c) = 0$ то $\mathrm{PMI}(t; c) = 0$.)

- Недостатък на поточкова мярка за взаимна информация е, че ако двете явления се срещнат само веднъж и то заедно, то мярката ще е много висока. Затова често се прилага праг, за да се избегнат редките явления.

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (25 мин)
3. Смятане с вектори и матрици (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разредено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
7. **Ранкиране на документи в контекстно пространство (5 мин)**

Влагане на документите в контекстно пространство

- Представянето на документите може да получим например като просто сумираме контекстните вектори съответстващи термовете, които се срещат в документа (BOW).
- Ранкирането може да се извърши според косинусовата близост спрямо вектора, получен за заявката.
- Съществуват значително по-релевантни методи за получаване на контекстния вектор съответстващ на документа — ще разглеждаме по-нататък в курса.
- Каква е гъстотата на векторите при това представяне?

Проблеми при влагането в контекстното пространство

- Размерността на контекстното пространство е броя на контекстите — може да бъде огромно.
- Даден терм може да се среща в стотици хиляди контексти. Документните вектори могат да съдържат милиони ненулеви елементи.
- Ранкирането в такова пространство е абсолютно невъзможно на практика.
- Следващата лекция ще разгледаме решение на този проблем