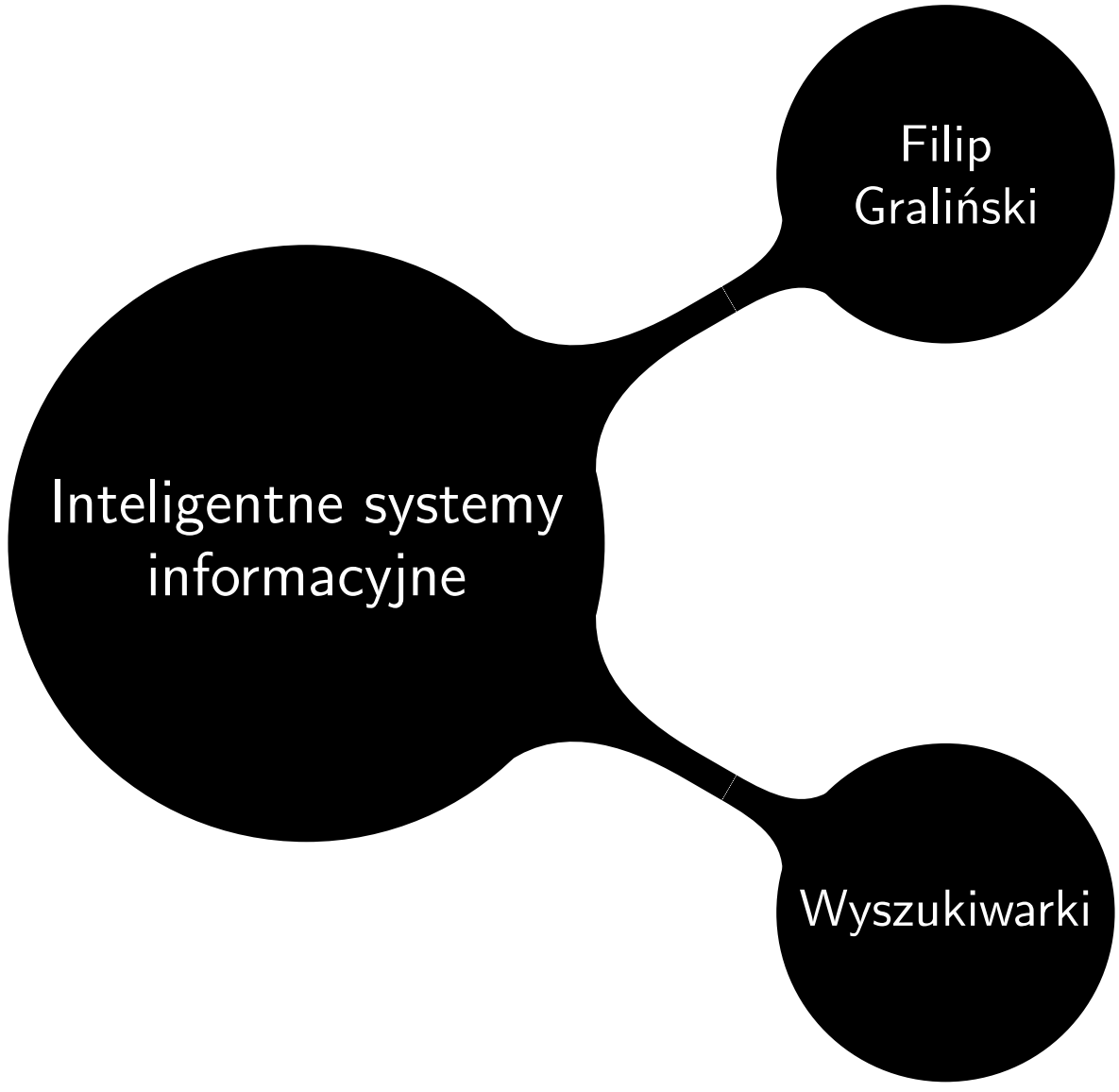
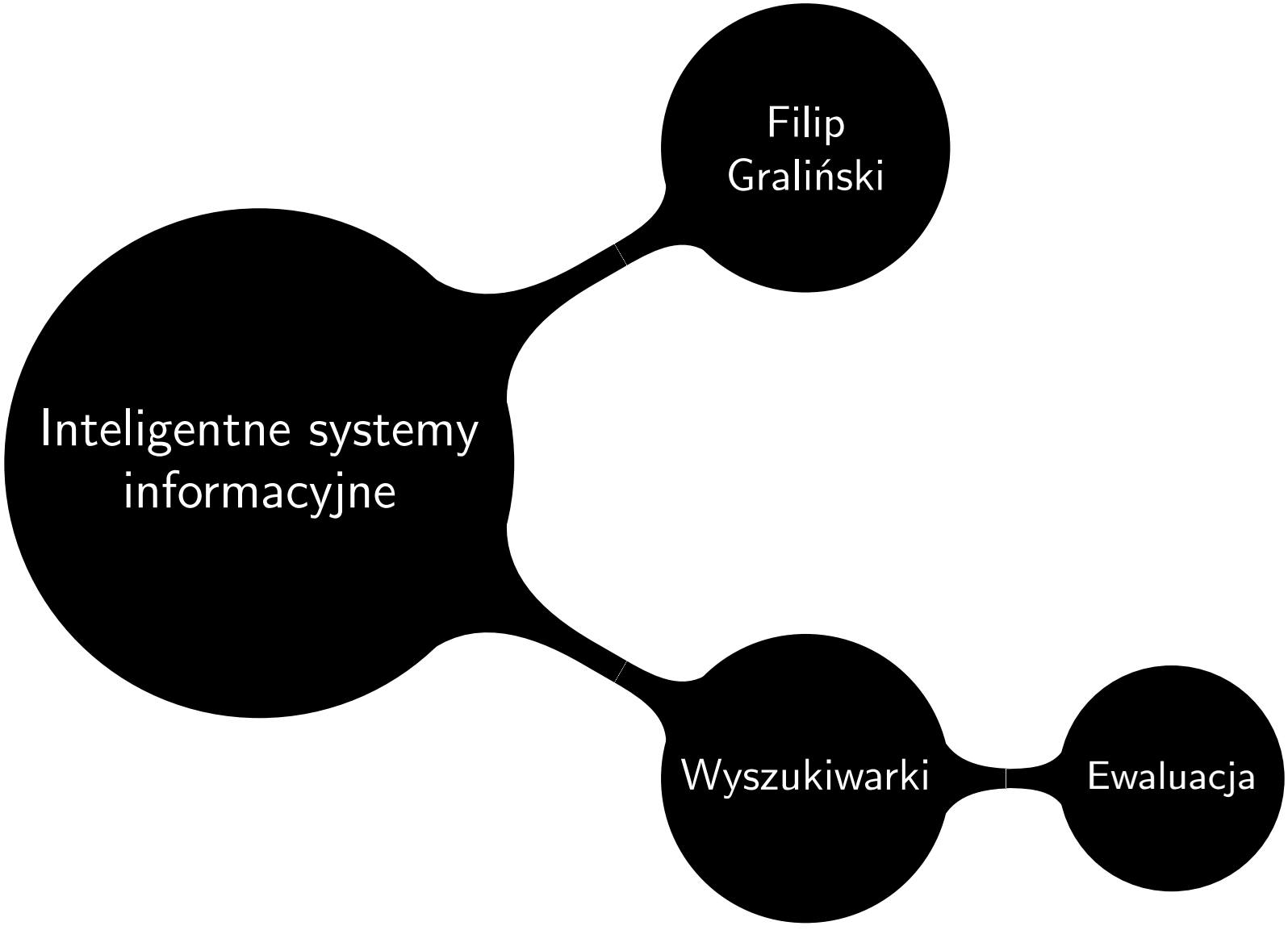

Filip
Graliński
Inteligentne systemy
informacyjne
Wyszukiwarki


Filip
Graliński
Inteligentne systemy
informacyjne
Wyszukiwarki
Ewaluacja

ogólnie o ocenianiu...

psychologiczne pułapki oceniania

- efekt pierwszeństwa
  błędy pojawiające się na początku wypracowania oceniane są surowiej niż błędy na końcu wypracowania

psychologiczne pułapki oceniania

- efekt pierwszeństwa
  błędy pojawiające się na początku wypracowania oceniane są surowiej niż błędy na końcu wypracowania
- efekt świeżości

psychologiczne pułapki oceniania

- efekt pierwszeństwa
  błędy pojawiające się na początku wypracowania oceniane są surowiej niż błędy na końcu wypracowania
- efekt świeżości
- efekt „aureoli”
  mamy skłonność do przypisywania innych pozytywnych cech ludziom atrakcyjnym fizycznie

psychologiczne pułapki oceniania

- efekt pierwszeństwa
  błędy pojawiające się na początku wypracowania oceniane są surowiej niż błędy na końcu wypracowania
- efekt świeżości
- efekt „aureoli”
  mamy skłonność do przypisywania innych pozytywnych cech ludziom atrakcyjnym fizycznie
- wpływ nastroju na ocenianie
  w pochmurne dni ludzie gorzej oceniają stan gospodarki niż w słoneczne

oceny korelacji

czy symptom świadczy o chorobie?

		choroba	
		obecna	nieobecna
symptom	obecny	37	33
	nieobecny	17	13

oceny korelacji

czy symptom świadczy o chorobie?

		choroba	
		obecna	nieobecna
symptom	obecny	37	33
	nieobecny	17	13

85% przepytanych pielęgniarek doszło do wniosku,
że istnieje związek między symptomem a chorobą

oceny korelacji

czy symptom świadczy o chorobie?

		choroba	
		obecna	nieobecna
symptom	obecny	37	33
	nieobecny	17	13

85% przepytanych pielęgniarek doszło do wniosku,
że istnieje związek między symptomem a chorobą

p-wartość przy teście χ^2: 0,8955

korelacja jest przez ludzi przeceniana,
regresja do średniej – niedoceniania

Jak uniknąć pułapek oceniania?

Jak uniknąć pułapek oceniania?

- stosować obiektywne miary

Jak uniknąć pułapek oceniania?

- stosować obiektywne miary
- najlepiej wymyślić syntetyczną funkcję oceny (pojedyncza liczba)

Jak uniknąć pułapek oceniania?

- stosować obiektywne miary
- najlepiej wymyślić syntetyczną funkcję oceny (pojedyncza liczba)
- jeśli to możliwe, stosować automatyczną ewaluację

ogólnie o ewaluacji...

Jakie własności ma dobra miara ewaluacji?

Jakie własności ma dobra miara ewaluacji?

- spełnia zdroworozsądkowe aksjomaty

Jakie własności ma dobra miara ewaluacji?

- spełnia zdroworozsądkowe aksjomaty
  - w szczególności dla przypadków brzegowych

Jakie własności ma dobra miara ewaluacji?

- spełnia zdroworozsądkowe aksjomaty
  - w szczególności dla przypadków brzegowych
- koreluje z ludzką oceną

Jakie własności ma dobra miara ewaluacji?

- spełnia zdroworozsądkowe aksjomaty
  - w szczególności dla przypadków brzegowych
- koreluje z ludzką oceną

Ewaluacja ewaluacji!

- np.: Joseph P. Turian, Luke Shen, I. Dan Melamed, *Evaluation of Machine Translation and its Evaluation*

Jak ewaluować wyszukiwarki?

Zacznijmy od elementarnych „cegiełek” ewaluacji: *precision*, *recall*.

miary ewaluacji – przykład

szukamy „byków”...

Należy pamiętać, że w sowim czasie Commodore 64 był najpopularniejszym mikrokomputerem. Sprzedan setki gier na ten komputer. Liczbę wyprodukowanych egzemplaży szacuje się na conajmniej 17 mln sztuk. Zmierzch komputera przypada na koniec lat 80. Commodore'a pokonały coraz potężniejsze pecety.

Wszystko to karze zastanowić się nad dziwnymi zakretami, jakimi szedł niepowstrzymany rozwój informatyki w ubiegłym wieku.

	rzeczywistość	
system	„byk”	nie-„byk”
„byk”		
nie-„byk”		

miary ewaluacji – przykład

szukamy „byków”...

Należy pamiętać, że w sowim czasie Commodore 64 był najpopularniejszym mikrokomputerem. Sprzedan setki gier na ten komputer. Liczbę wyprodukowanych egzemplaży szacuje się na conajmniej 17 mln sztuk. Zmierzch komputera przypada na koniec lat 80. Commodore'a pokonały coraz potężniejsze pecety.

Wszystko to karze zastanowić się nad dziwnymi zakretami, jakimi szedł niepowstrzymany rozwój informatyki w ubiegłym wieku.

	rzeczywistość	
system	„byk”	nie-„byk”
„byk”		
nie-„byk”		

miary ewaluacji – przykład

szukamy „byków”...

Należy pamiętać, że w sowim czasie Commodore 64 był najpopularniejszym mikrokomputerem. Sprzedan setki gier na ten komputer. Liczbę wyprodukowanych egzemplaży szacuje się na conajmniej 17 mln sztuk. Zmierzch komputera przypada na koniec lat 80. Commodore'a pokonały coraz potężniejsze pecety.

Wszystko to karze zastanowić się nad dziwnymi zakretami, jakimi szedł niepowstrzymany rozwój informatyki w ubiegłym wieku.

	rzeczywistość	
system	„byk”	nie-„byk”
„byk”	5 *true positives*	
nie-„byk”		

miary ewaluacji – przykład

szukamy „byków”...

Należy pamiętać, że w sowim czasie Commodore 64 był najpopularniejszym mikrokomputerem. Sprzedan setki gier na ten komputer. Liczbę wyprodukowanych egzemplaży szacuje się na conajmniej 17 mln sztuk. Zmierzch komputera przypada na koniec lat 80. Commodore'a pokonały coraz potężniejsze pecety.

Wszystko to karze zastanowić się nad dziwnymi zakretami, jakimi szedł niepowstrzymany rozwój informatyki w ubiegłym wieku.

	rzeczywistość	
system	„byk”	nie-„byk”
„byk”	5 *true positives*	3 *false positives*
nie-„byk”		

miary ewaluacji – przykład

szukamy „byków”...

Należy pamiętać, że w sowim czasie Commodore 64 był najpopularniejszym mikrokomputerem. Sprzedan setki gier na ten komputer. Liczbę wyprodukowanych egzemplaży szacuje się na conajmniej 17 mln sztuk. Zmierzch komputera przypada na koniec lat 80. Commodore'a pokonały coraz potężniejsze pecety.

Wszystko to karze zastanowić się nad dziwnymi zakretami, jakimi szedł niepowstrzymany rozwój informatyki w ubiegłym wieku.

	rzeczywistość	
system	„byk”	nie-„byk”
„byk”	5 *true positives*	3 *false positives*
nie-„byk”	2 *false negatives*	

miary ewaluacji – przykład

szukamy „byków”...

Należy pamiętać, że w sowim czasie Commodore 64 był najpopularniejszym mikrokomputerem. Sprzedan setki gier na ten komputer. Liczbę wyprodukowanych egzemplaży szacuje się na conajmniej 17 mln sztuk. Zmierzch komputera przypada na koniec lat 80. Commodore'a pokonały coraz potężniejsze pecety.

Wszystko to karze zastanowić się nad dziwnymi zakretami, jakimi szedł niepowstrzymany rozwój informatyki w ubiegłym wieku.

	rzeczywistość	
system	„byk”	nie-„byk”
„byk”	5 *true positives*	3 *false positives*
nie-„byk”	2 *false negatives*	44 *true negatives*

precision i *recall*

Definicja

Precision („precyzja”) – jaki odsetek jednostek wskazanych przez system to faktycznie „byki”

$$P = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}}$$

precision i *recall*

Definicja

Precision („precyzja”) – jaki odsetek jednostek wskazanych przez system to faktycznie „byki”

$$P = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}}$$

Definicja

Recall („kompletność”) – jaki odsetek „byków” został wskazany przez system

$$R = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}}$$

przykład cd.

	rzeczywistość	
system	„byk”	nie-„byk”
„byk”	5 *true positives*	3 *false positives*
nie-„byk”	2 *false negatives*	44 *true negatives*

$$P = \frac{5}{5+3} = 0{,}625$$

przykład cd.

system	*rzeczywistość* „byk”	*rzeczywistość* nie-„byk”
„byk”	5 *true positives*	3 *false positives*
nie-„byk”	2 *false negatives*	44 *true negatives*

$$P = \frac{5}{5+3} = 0{,}625 \quad R = \frac{5}{5+2} \approx 0{,}714$$

własności *precision* i *recall*

- $P, R \in [0, 1]$

własności *precision* i *recall*

- $P, R \in [0, 1]$
- idealny system: $P = 1{,}0$, $R = 1{,}0$

własności *precision* i *recall*

- $P, R \in [0, 1]$
- idealny system: $P = 1{,}0$, $R = 1{,}0$
- głupi system nadgorliwy: $P \approx 0{,}0$, $R = 1{,}0$

własności *precision* i *recall*

- $P, R \in [0, 1]$
- idealny system: $P = 1{,}0$, $R = 1{,}0$
- głupi system nadgorliwy: $P \approx 0{,}0$, $R = 1{,}0$
- głupi system leniwy: $P = 1{,}0$, $R = 0{,}0$

własności *precision* i *recall*

- $P, R \in [0, 1]$
- idealny system: $P = 1{,}0$, $R = 1{,}0$
- głupi system nadgorliwy: $P \approx 0{,}0$, $R = 1{,}0$
- głupi system leniwy: $P = 1{,}0$, $R = 0{,}0$

Za cenę spadku precyzji można zwiększyć kompletność i *vice versa*

dlaczego nie *accuracy* / *error*?

Definicja

Accuracy („dokładność”) – jaki odsetek poprawnych odpowiedzi systemu

$$A = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn}}$$

dlaczego nie *accuracy*/*error*?

Definicja

Accuracy („dokładność") – jaki odsetek poprawnych odpowiedzi systemu

$$A = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn}}$$

Definicja

Error („błąd") – jaki odsetek złych odpowiedzi systemu

$$E = \frac{\mathrm{fp} + \mathrm{fn}}{\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn}} = 1 - A$$

przykład cd.

system	*rzeczywistość* „byk”	*rzeczywistość* nie-„byk”
„byk”	5 *true positives*	3 *false positives*
nie-„byk”	2 *false negatives*	44 *true negatives*

$A = \frac{5+44}{5+44+2+3} \approx 0{,}907$

przykład cd.

system	*rzeczywistość* „byk”	*rzeczywistość* nie-„byk”
„byk”	5 *true positives*	3 *false positives*
nie-„byk”	2 *false negatives*	44 *true negatives*

$$A = \frac{5+44}{5+44+2+3} \approx 0{,}907 \qquad E = 1 - A \approx 0{,}093$$

przykład cd.

	rzeczywistość	
system	„byk”	nie-„byk”
„byk”	5 *true positives*	3 *false positives*
nie-„byk”	2 *false negatives*	44 *true negatives*

$$A = \frac{5+44}{5+44+2+3} \approx 0{,}907$$

Duże *accuracy* może nic znaczyć w wypadku systemów, które wykrywają coś rzadkiego!

miara F

Miara F pozwala za pomocą jednej liczby ująć kombinację P i R

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$

współczynnik $\alpha \in [0, 1]$ pozwala w różny sposób „ważyć” P i R

miara F

Miara F pozwala za pomocą jednej liczby ująć kombinację P i R

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

współczynnik $\alpha \in [0, 1]$ pozwala w różny sposób „ważyć” P i R

- jeśli $\alpha = 0{,}5$, to $F = \frac{2PR}{P+R}$

miara F

Miara F pozwala za pomocą jednej liczby ująć kombinację P i R

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

współczynnik $\alpha \in [0,1]$ pozwala w różny sposób „ważyć” P i R

- jeśli $\alpha = 0{,}5$, to $F = \frac{2PR}{P+R}$
- jeśli $\alpha = 1{,}0$, to $F = P$

miara F

Miara F pozwala za pomocą jednej liczby ująć kombinację P i R

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

współczynnik $\alpha \in [0,1]$ pozwala w różny sposób „ważyć” P i R

- jeśli $\alpha = 0{,}5$, to $F = \frac{2PR}{P+R}$
- jeśli $\alpha = 1{,}0$, to $F = P$
- jeśli $\alpha = 0{,}0$, to $F = R$

miara F

Miara F pozwala za pomocą jednej liczby ująć kombinację P i R

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

współczynnik $\alpha \in [0,1]$ pozwala w różny sposób „ważyć” P i R

- jeśli $\alpha = 0{,}5$, to $F = \frac{2PR}{P+R}$
- jeśli $\alpha = 1{,}0$, to $F = P$
- jeśli $\alpha = 0{,}0$, to $F = R$

Dlaczego średnia harmoniczna, a nie arytmetyczna?

miara F

Miara F pozwala za pomocą jednej liczby ująć kombinację P i R

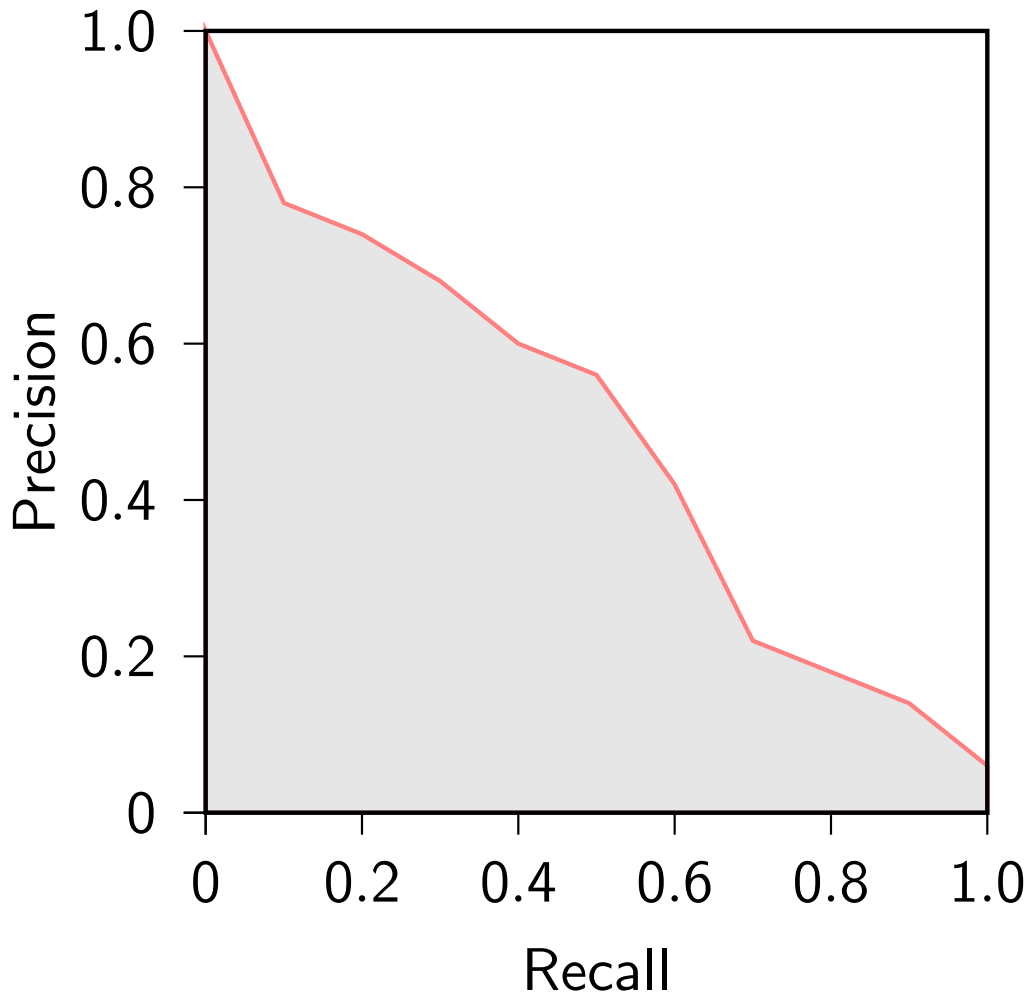
$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

współczynnik $\alpha \in [0,1]$ pozwala w różny sposób „ważyć” P i R

- jeśli $\alpha = 0{,}5$, to $F = \frac{2PR}{P+R}$
- jeśli $\alpha = 1{,}0$, to $F = P$
- jeśli $\alpha = 0{,}0$, to $F = R$

Dlaczego średnia harmoniczna, a nie arytmetyczna? Bo system z małym recallem, mógłby mieć $F \approx 0{,}5$

krzywa *precision-recall*



gdzie *precision* i *recall*?

- wyszukiwanie informacji (*information retrieval*)
- korekta ortografii i gramatyki

gdzie *precision* i *recall*?

- wyszukiwanie informacji (*information retrieval*)
- korekta ortografii i gramatyki
- filtry antyspamowe
- systemy IDS (*Intrusion Detection System*)
- ...

o ewaluacji wyszukiwarek...

Która wyszukiwarka jest lepsza?

✔	1.	____
✗	2.	____
✗	3.	____
✗	4.	____
✗	5.	____
✗	6.	____
✗	7.	____
✔	8.	____
✗	9.	____
✗	10.	____

✗	1.	____
✗	2.	____
✔	3.	____
✔	4.	____
✔	5.	____
✗	6.	____
✗	7.	____
✗	8.	____
✗	9.	____
✗	10.	____

Zakładamy binarną ocenę poszczególnych wyników:

- ✔– wynik relewantny (*relevant*)
- ✗– wynik nierelewantny (*irrelevant*)

Która wyszukiwarka jest lepsza?

✔	1.	____
✗	2.	____
✗	3.	____
✗	4.	____
✗	5.	____
✗	6.	____
✗	7.	____
✔	8.	____
✗	9.	____
✗	10.	____

✗	1.	____
✗	2.	____
✔	3.	____
✔	4.	____
✔	5.	____
✗	6.	____
✗	7.	____
✗	8.	____
✗	9.	____
✗	10.	____

Zakładamy binarną ocenę poszczególnych wyników:

- ✔– wynik relewantny (*relevant*)
- ✗– wynik nierelewantny (*irrelevant*)

Ocena względem potrzeby informacyjnej, nie zapytania.

Zacznijmy od czegoś prostego...

✔	1.	____
✗	2.	____
✗	3.	____
✗	4.	____
✗	5.	____
✗	6.	____
✗	7.	____
✔	8.	____
✗	9.	____
✗	10.	____

✗	1.	____
✗	2.	____
✔	3.	____
✔	4.	____
✔	5.	____
✗	6.	____
✗	7.	____
✗	8.	____
✗	9.	____
✗	10.	____

(Zakładamy, że wszystkich relewantnych wyników jest 6)

Zacznijmy od czegoś prostego...

✔	1.	____
✗	2.	____
✗	3.	____
✗	4.	____
✗	5.	____
✗	6.	____
✗	7.	____
✔	8.	____
✗	9.	____
✗	10.	____

✗	1.	____
✗	2.	____
✔	3.	____
✔	4.	____
✔	5.	____
✗	6.	____
✗	7.	____
✗	8.	____
✗	9.	____
✗	10.	____

(Zakładamy, że wszystkich relewantnych wyników jest 6)

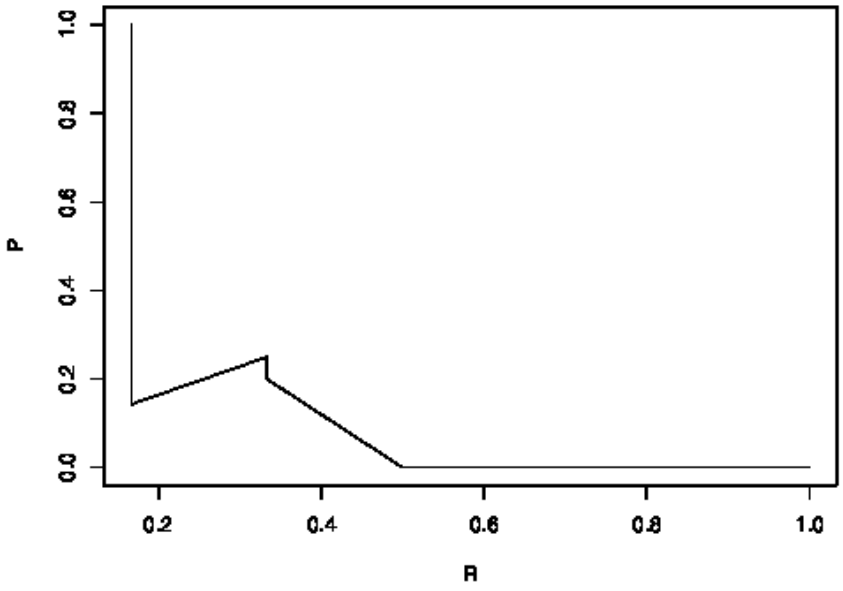
$P = 0{,}2$
$R = 0{,}333$
$F = 0{,}25$

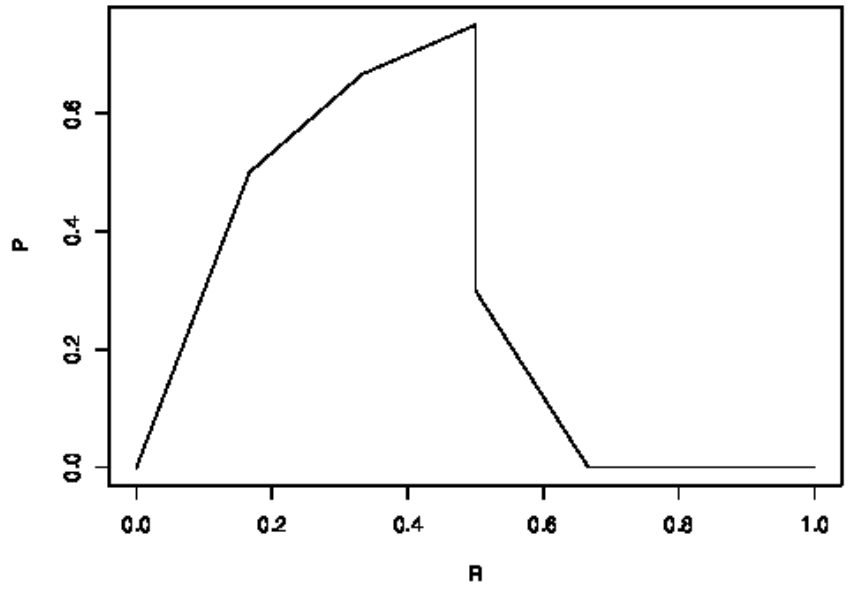
$P = 0{,}3$
$R = 0{,}5$
$F = 0{,}375$

W ogóle nie bierzemy pod uwagę kolejności!

Jak wziąć pod uwagę kolejność?

Krzywa *precision-recall*



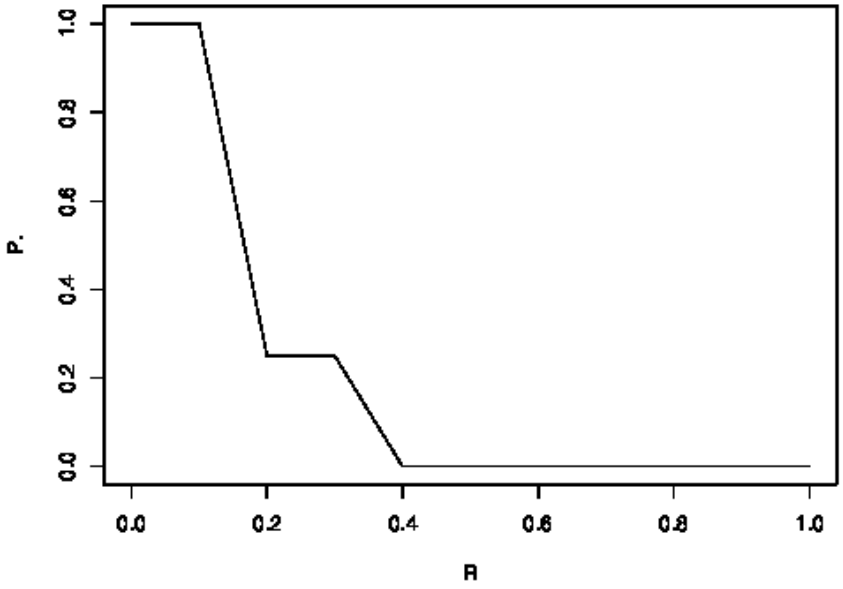


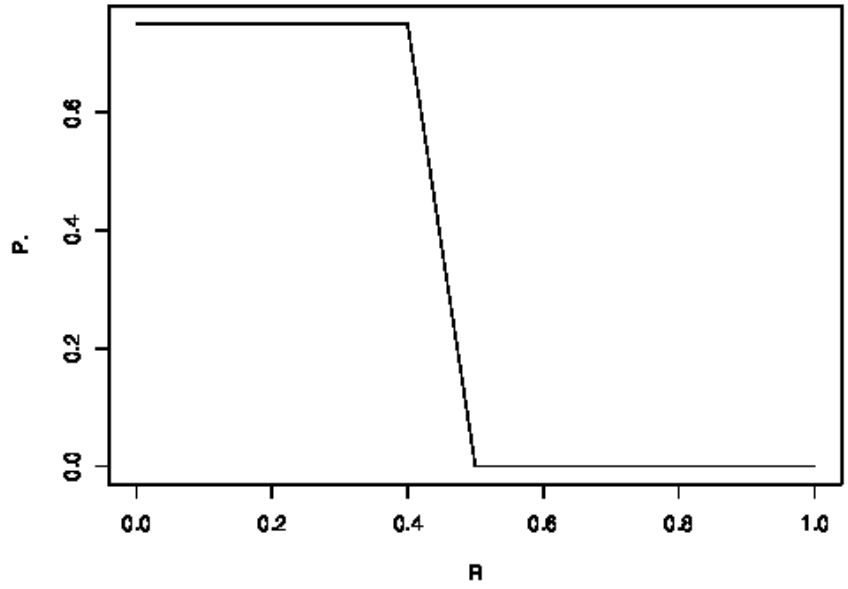
Interpolowana *precision*

$$p^*(r) = \max_{r' \geqslant r} p(r')$$

- – unikamy „zębów” na wykresie
- – może być wyznaczona dla dowolnej wartości r (także dla 0)

11-point interpolated average precision





Tak naprawdę wyliczana nie dla pojedynczego zapytania, tylko uśredniana dla każdego z 11 punktów na całym zbiorze testowym!

Ale chcemy pojedynczą liczbę?!?

$$MAP = Mean\ Average\ Precision$$

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} p(R_{jk})$$

- Q – zbiór potrzeb informacyjnych
- m_j – liczba relewantnych dokumentów dla potrzeby $q_j \in Q$
- R_{jk} – zbiór wyników aż do dokumentu d_k dla potrzeby q_j

Nie tylko relewantność

Ostatecznie chcemy ewaluować zadowolenie użytkownika:

- szybkość działania
- jakość zajawek
- „ładność” serwisu
- ...

Nie tylko relewantność

Ostatecznie chcemy ewaluować zadowolenie użytkownika:

- szybkość działania
- jakość zajawek
- „ładność” serwisu
- ...

...no i oczywiście pogoda za oknem

Bibliografia

T. Tyszka.
Psychologiczne pułapki oceniania i podejmowania decyzji.
GWP, 1999.

Bibliografia

- T. Tyszka.
  Psychologiczne pułapki oceniania i podejmowania decyzji.
  GWP, 1999.
- S. Sutherland.
  Rozum na manowcach. Dlaczego postępujemy irracjonalnie?
  Książka i Wiedza, 1996.

Bibliografia


T. Tyszka.
Psychologiczne pułapki oceniania i podejmowania decyzji.
GWP, 1999.


S. Sutherland.
Rozum na manowcach. Dlaczego postępujemy irracjonalnie?
Książka i Wiedza, 1996.


Ch. D. Manning i H. Schütze.
Foundations of Statistical Natural Language Processing
The MIT Press, 2003.

Bibliografia


T. Tyszka.
Psychologiczne pułapki oceniania i podejmowania decyzji.
GWP, 1999.


S. Sutherland.
Rozum na manowcach. Dlaczego postępujemy irracjonalnie?
Książka i Wiedza, 1996.


Ch. D. Manning i H. Schütze.
Foundations of Statistical Natural Language Processing
The MIT Press, 2003.


A. Farghaly (red.)
Handbook for Language Engineers
CSLI Publications, 2003.

Bibliografia


T. Tyszka.
Psychologiczne pułapki oceniania i podejmowania decyzji.
GWP, 1999.


S. Sutherland.
Rozum na manowcach. Dlaczego postępujemy irracjonalnie?
Książka i Wiedza, 1996.


Ch. D. Manning i H. Schütze.
Foundations of Statistical Natural Language Processing
The MIT Press, 2003.


A. Farghaly (red.)
Handbook for Language Engineers
CSLI Publications, 2003.