

# Avtomatsko ocenjevanje esejev s sistemom SAGE+

---

Kaja Zupanc

Jota, 25. 5. 2017

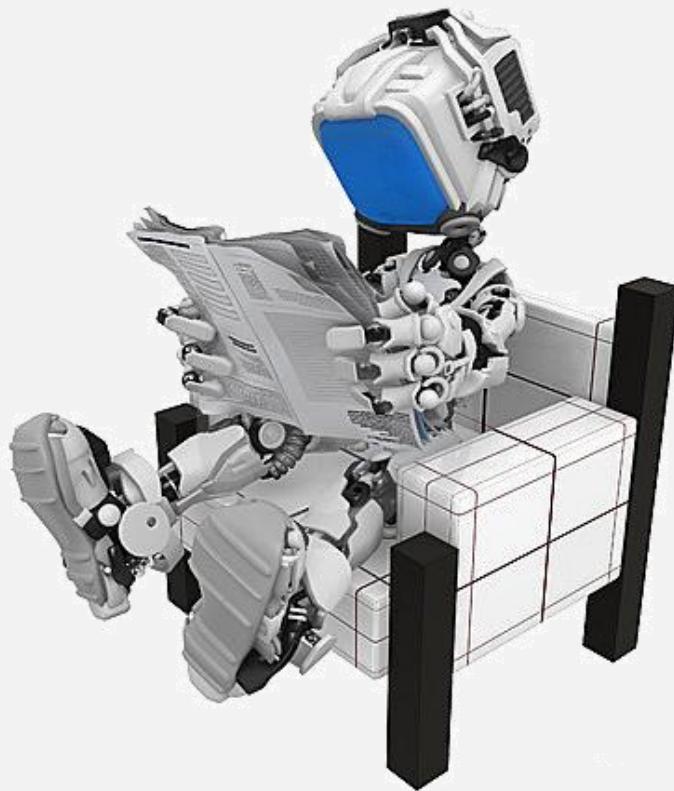
[kaja.zupanc@fri.uni-lj.si](mailto:kaja.zupanc@fri.uni-lj.si)

Univerza v Ljubljani  
Fakulteta za računalništvo  
*in informatiko*

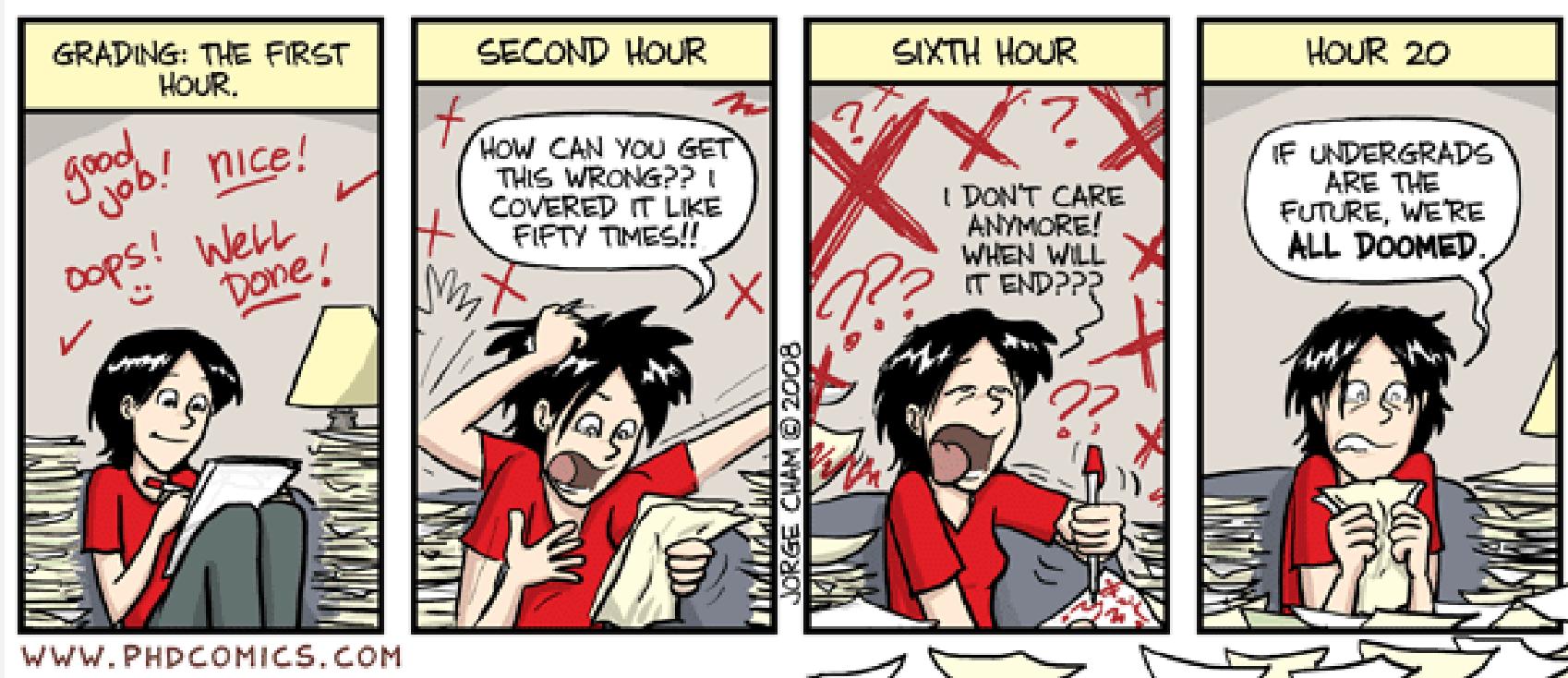


# Oris vsebine

1. Predstavitev problema.
2. Razvoj področja.
3. Predstavitev sistema SAGE+.



# Problem?



# Avtomatsko ocenjevanje esejev (AEE)

## Definicija:

Proces ovrednotenja in ocenjevanja zapisanega besedila z računalniškim programom.<sup>1</sup>



1. M. D. Shermis and J. Burstein, "Introduction," in *Automated essay scoring: A cross-disciplinary perspective*, M. D. Shermis and J. Burstein, Eds. Manwah, NJ: Lawrence Erlbaum Associates, 2003, xiii–xvi.

# Razvoj

Prvi sistem predstavljen že leta 1966 in dostopen leta 1973<sup>1</sup>.

**Pro:** čas, cena, objektivnost, razlike med ocenjevalci  
&

**Contra:** ni človeškega razumevanja

Hitrejši razvoj se prične v 90ih letih.

Danes se v tujini AEE sistemi uporablajo:

- V **kombinaciji s človeškim ocenjevalcem** pri standardiziranih testih, kot so Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL), Graduate Management Admissions Test (GMAT), American College Testing (ACT), ...
- Kot **edini ocenjevalec** v sistemih za vajo ali nekaterih manjših preizkusih.

1. H. B. Ajay, P. I. Tillet, and E. B. Page, "Analysis of essays by computer (AEC-II)," U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development, Washington, D.C., Tech. Rep., 1973.

# State-of-the-art sistemi

Glavna ovira razvoja: sistemi so komercialni in niso javno dostopni.

Komercialni sistemi:

- Project Essay Grade<sup>1</sup>,
- E-rater<sup>2</sup>,
- Intelligent Essay Assessor<sup>3</sup>,
- IntelliMetric<sup>4</sup>.

Pomanjkljivost: ocenjevanje semantike besedila in povratne informacije o vsebini.

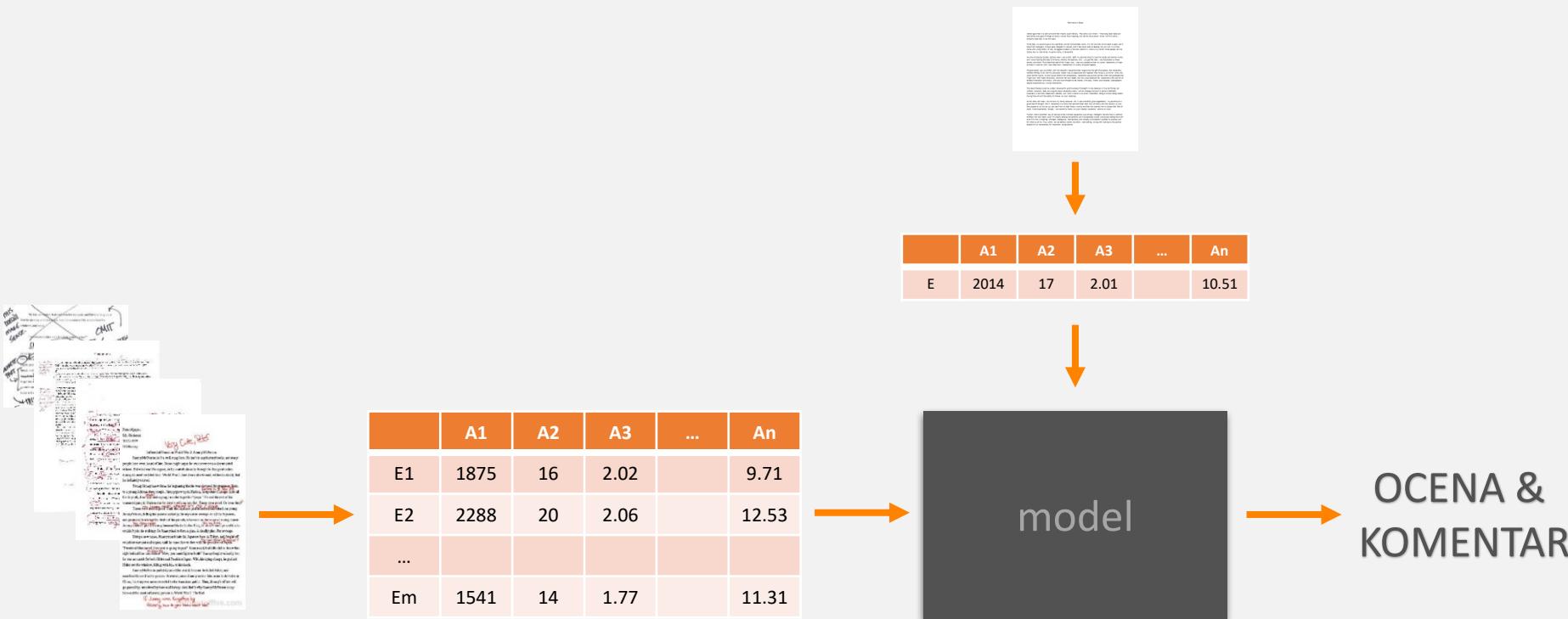
1. M. D. Shermis, C. M. Koch, E. B. Page, T. Z. Keith, S. Harrington, "Trait Ratings for Automated Essay Grading," *Educ. Psychol. Meas.*, vol. 62, no. 5, pp. 5–18, Feb. 2002

2. J. Burstein, J. Tetreault, N. Madnani, "The E-rater® Automated Essay Scoring System," in *Handbook of AEE*, New York: Routledge, 2013, pp. 55–67.

3. P. W. Foltz, L. A. Streeter, K. E. Lochbaum, T. K. Landauer, "Implementation and Applications of the Intelligent Essay Assessor," in *Handbook of AEE*, New York: Routledge, 2013, pp. 68–88.

4. M. T. Schultz, "The IntelliMetric Automated Essay Scoring Engine," in *Handbook of AEE*, New York: Routledge, 2013, pp. 89–98.

# Avtomatsko ocenjevanje esejev



# SAGE+: Semantic Automated Grader for Essays

Atributi:

- sintaksa (leksikalna dovršenost, slovnica, mehanika) – 67,
- vsebina (primerjava z ostalimi eseji) – 5,
- skladnost (povezanost besedila) – 29 + 32,
- semantika (semantične napake) – 3.

# SAGE+: Semantic Automated Grader for Essays

Atributi:

1. sintaksa (leksikalna dovršenost, slovica, mehanika) – 67,  
vsebina (primerjava z ostalimi eseji) – 5,
2. skladnost (povezanost besedila) – 29 + 32,
3. semantika (semantične napake) – 3.

# 1. skupina atributov: sintaksa & vsebina

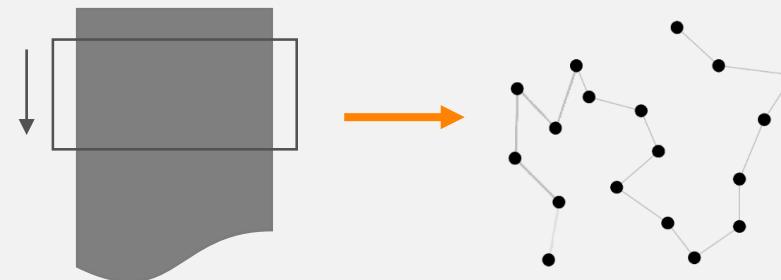
Content	Lexical sophistication	Grammar	Mechanics
<p>cosine similarity with source text, score point level for maximum cosine similarity over all score points,</p> <p>cosine similarity with essays that have highest score point level,</p> <p>pattern cosine,</p> <p>weighted sum of all cosine correlation values,</p>	<p>number of characters, number of words, most frequent word length, average word length, number of long words, number of short words, number of sentences, number of long sentences, number of short sentences, most frequent sentence length, average sentence length, number of different words, number of stopwords, Readability measures Gunning Fog index, Flesch reading ease, Flesch Kincaid grade level, Dale-Chall readability formula, automated readability index, simple measure of Gobbledygook, word variation index, LIX, nominal ratio, Lexical diversity type token ration, Guiraud's index, Yule's K, the D estimate, hapax legomena, advanced Guiraud,</p>	<p>number of different POS tags, height of the sentence structure tree, correct verb form, number of grammar errors, Number of each POS tag preposition/subordinating conjunction, coordinating conjunction, comparative adjective, superlative adjective, singular or mass common noun, comparative adjective, plural common noun, singular proper noun, plural proper noun, verb - gerund/present participle, verb - past participle, verb - 3rd person sing. present, possessive wh-pronoun, comparative adverb, superlative adverb, verb - base form, particle, predeterminer, preposition, personal pronoun, adjective, numeral, existential there, wh-determiner, wh-pronoun,</p> <p>verb - past tense, modal auxiliary, genitive marker, participle, possesive pronoun, adverb, determiner, modal auxiliary, wh-determiner, wh-adverb,</p>	<p>number of spellchecking errors, number of capitalization errors, number of punctuation errors.</p>

## 2. skupina atributov: skladnost

Atributi skladnosti:

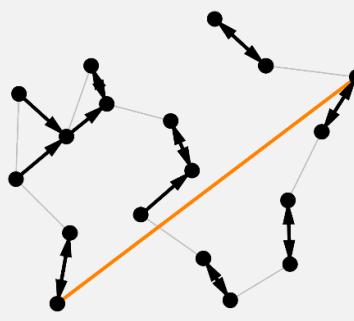
- prostorska skladnost – 29,
- analiza omrežij – 35.

$$TF - IDF(t, p, d, D) = TF(t, d) \cdot IDF(t, d) = \frac{|\{w \in p : w = t\}|}{|\{w \in d\}|} \cdot \log \frac{|\{d \in D\}|}{|\{d \in D : t \in d\}|}$$



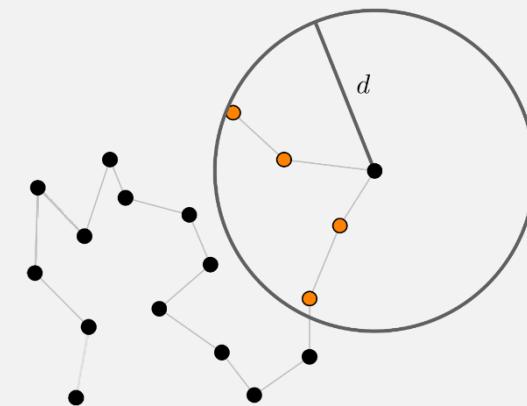
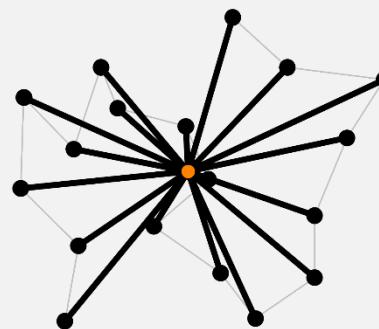
# Prostorska skladnost

1. Osnovne mere skladnosti besedila.
2. Prostorska skladnost.
3. Prostorska avtokorelacija.



$$R = \frac{\sum_{i=1}^N r_i}{\frac{1}{2\sqrt{N}}}$$

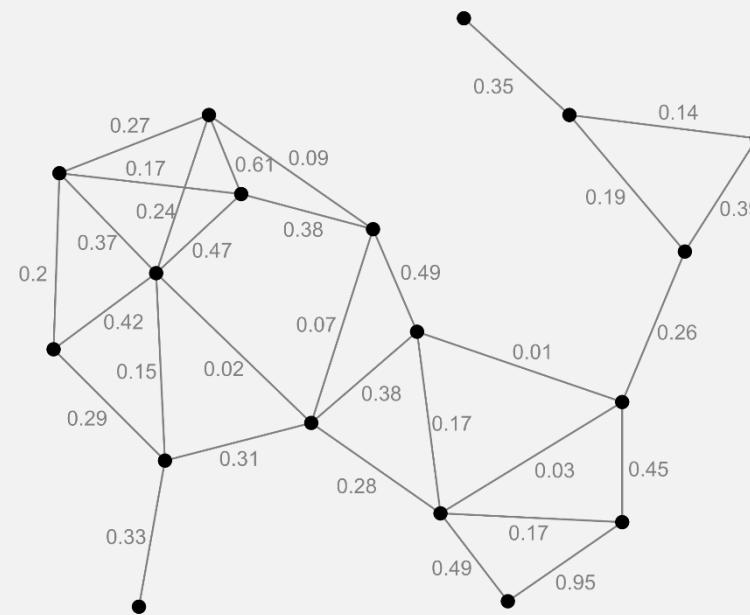
$$S_D = \sqrt{\frac{\sum_{k=1}^n \sum_{i=1}^N (D_i^k - \bar{D}_c^k)^2}{N}}$$



$$G(d) = \frac{1}{n} \sum_{k=1}^n \left[ \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij}(d) D_i^k D_j^k}{\sum_{i=1}^N \sum_{j=1}^N D_i^k D_j^k} \right]$$

# Analiza omrežij

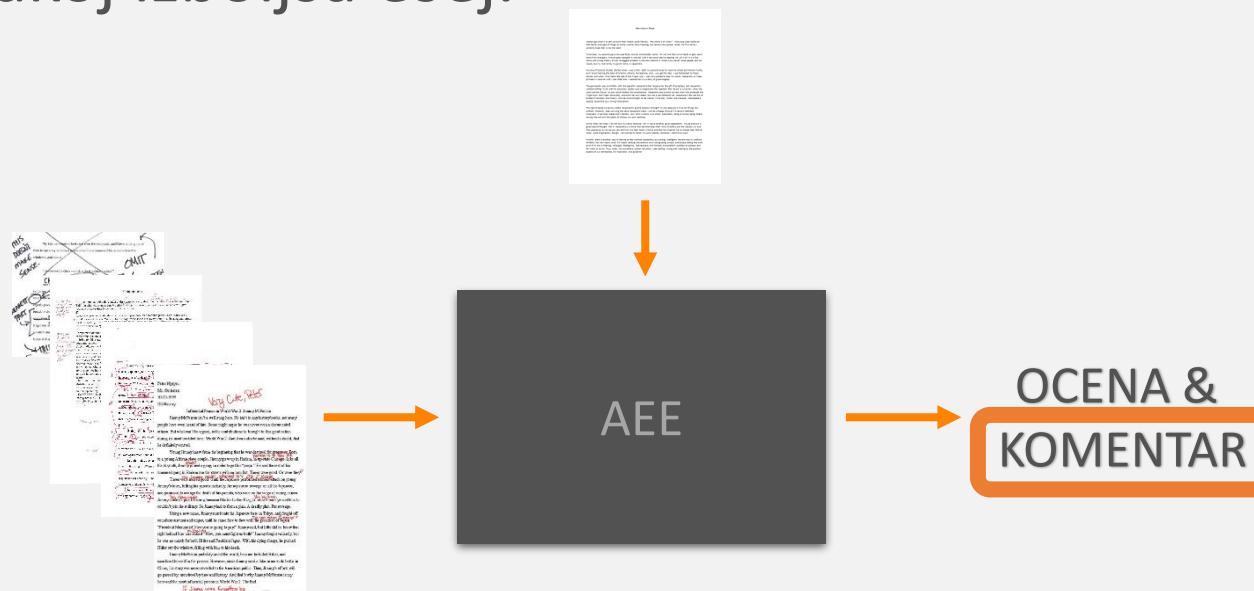
- Omrežje stavčne podobnosti (ang.: sentence-similarity network).
- Vsak esej predstavlja svoje omrežje.
- Atributi:
  1. Osnovne strukturne metrike.
  2. Sestavljene strukturne metrike.
  3. Metrike omrežne entropije.



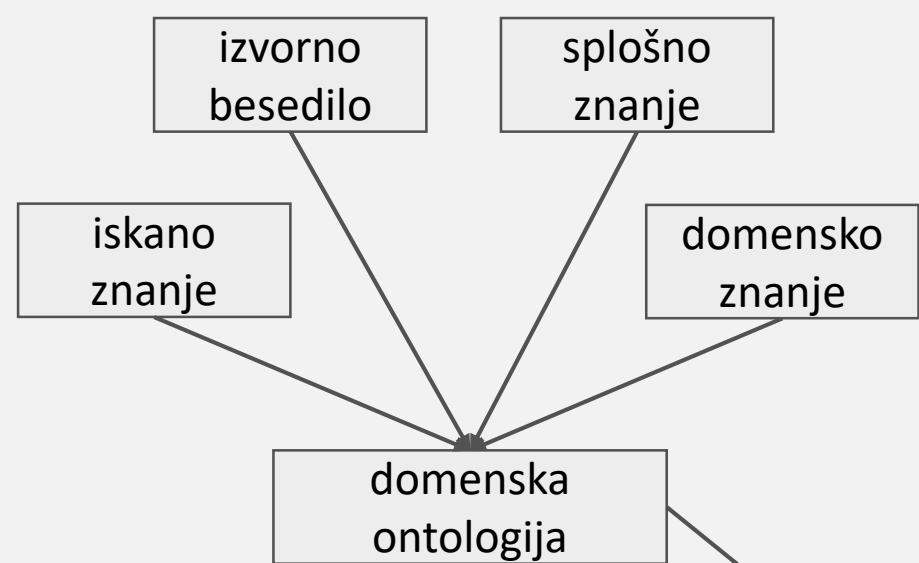
# Je ocena res vse?

Standardizirani testi → Napoved končne ocene.

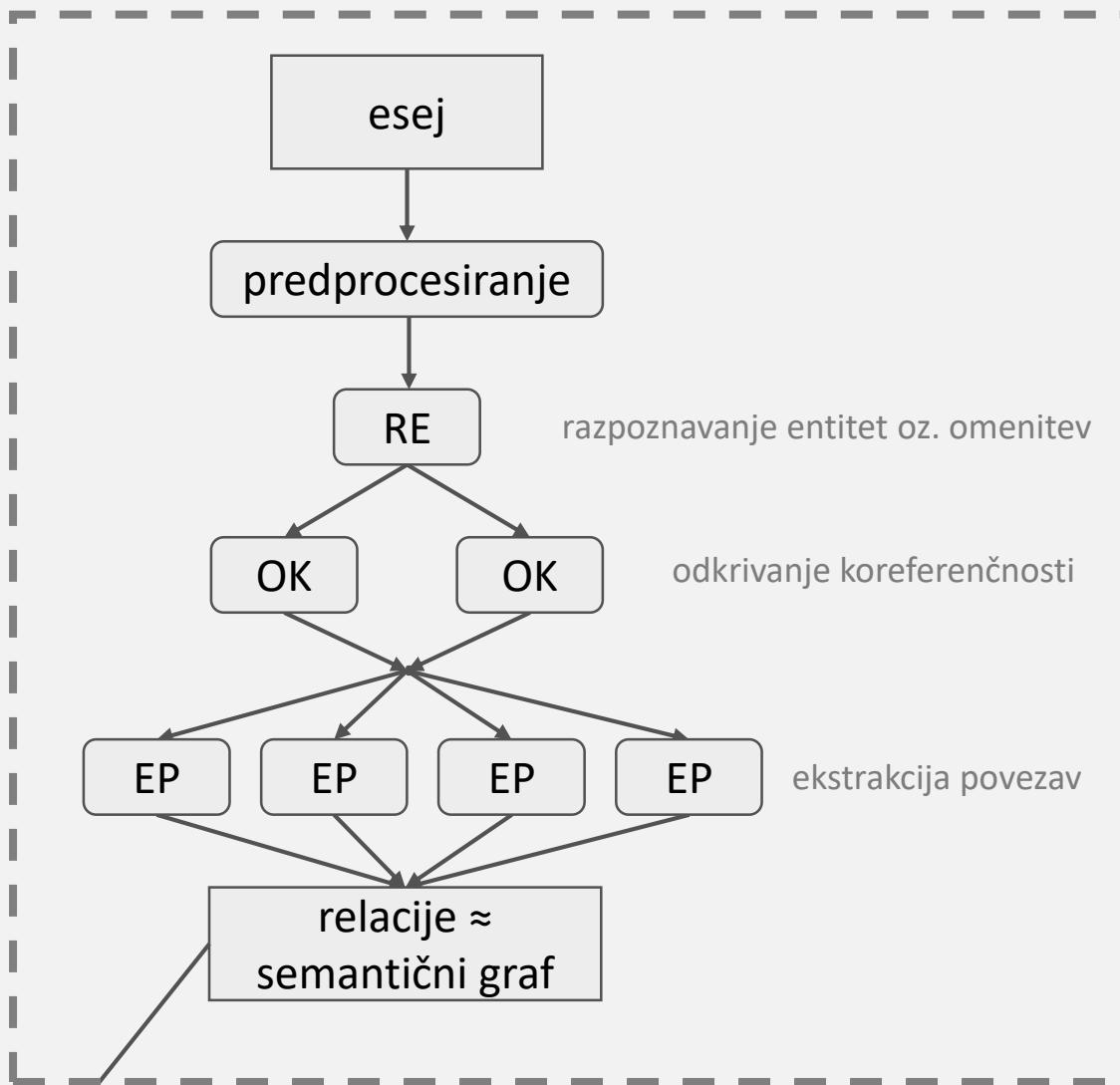
Pripomoček pri učenju → Učencu nudi takojšnjo povratno informacijo, učenec lahko takoj izboljša esej.



# Sistem za avtomatsko detekcijo semantičnih napak



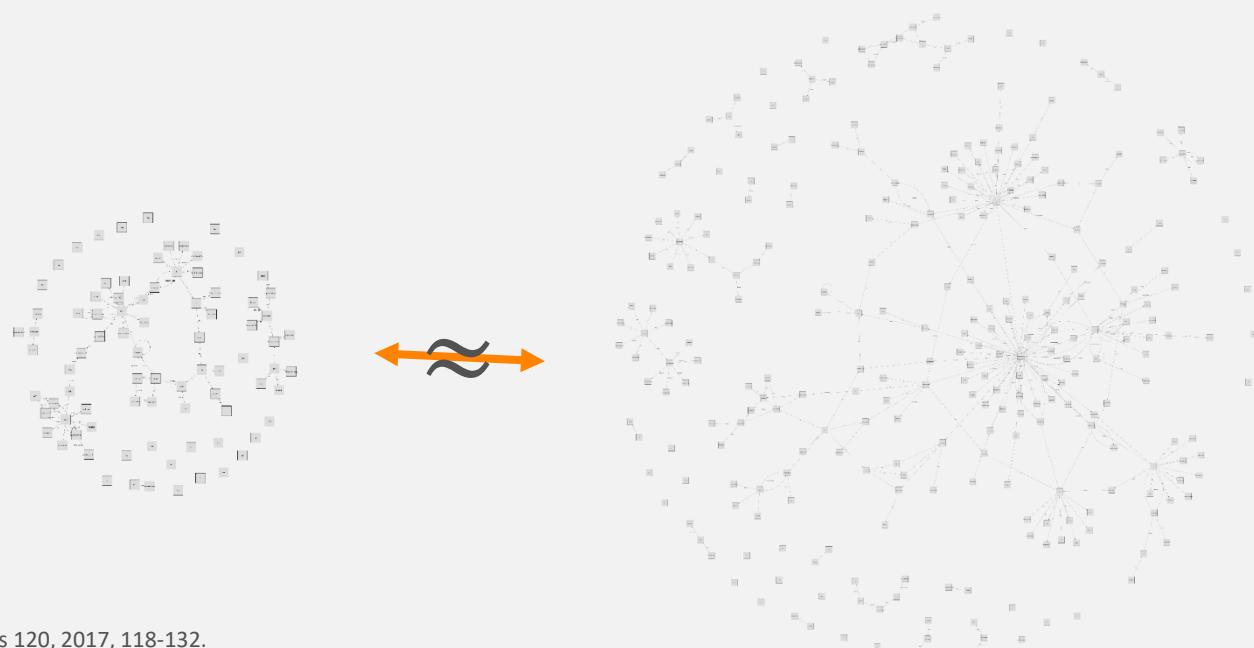
Hermit logic reasoner<sup>1</sup>



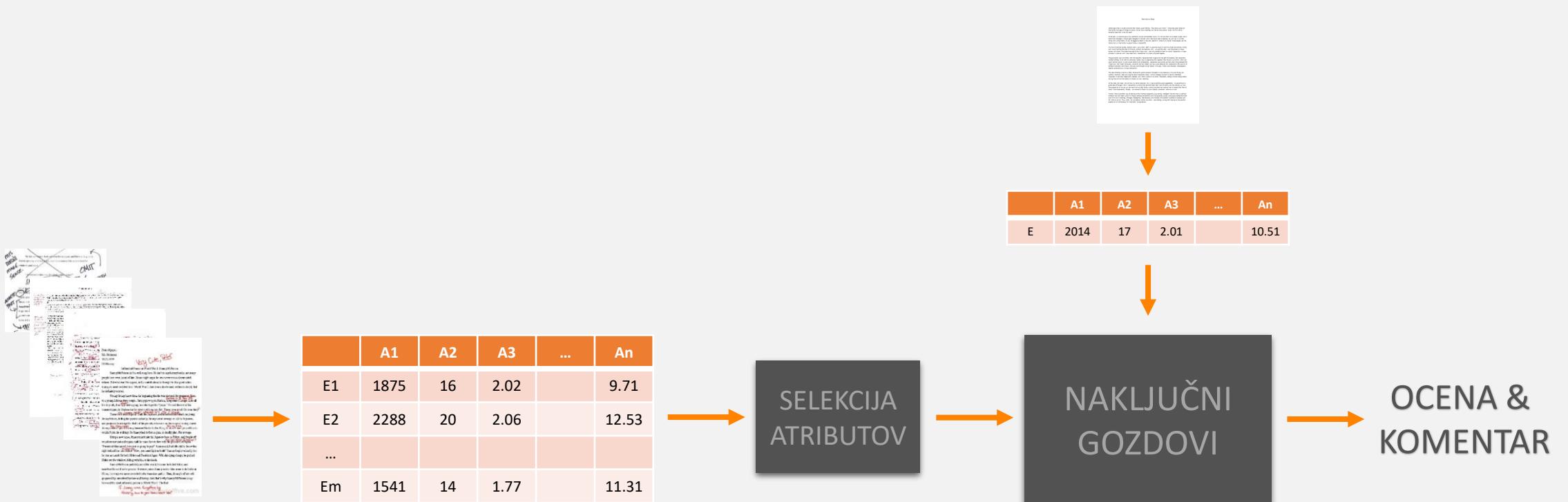
<sup>1</sup>B. Motik , R. Shearer , I. Horrocks , Hypertableau reasoning for description logics, J. Artif. Intell. Res. 36 (2009) 165–228 .

### 3. Atributi: semantika

1. Število nezadovoljivih pimerov, ko dodajamo razrede in objekte v ontologijo.
2. Število nekonsistentnih napak, ko dodajamo relacije v ontologijo,
3. Celotno število konsistentnih napak (vsota prvih dveh atributov).



# SAGE+: Semantic Automated Grader for Essays



E	A1	A2	A3	...	An
E1	1875	16	2.02		9.71
E2	2288	20	2.06		12.53
...					
Em	1541	14	1.77		11.31

SELEKCIJA  
ATRIBUTOV

NAKLJUČNI  
GOZDOVI

OCENA &  
KOMENTAR



E	A1	A2	A3	...	An
E	2014	17	2.01		10.51
...					



# Rezultati

Sistem	DS1	DS2a	DS2b	DS3	DS4	DS5	DS6	DS7	DS8
SAGE+	<b>0.93</b>	<b>0.79</b>	<b>0.67</b>	<b>0.83</b>	<b>0.81</b>	<b>0.89</b>	<b>0.79</b>	<b>0.88</b>	<b>0.81</b>
PEG	0.82	0.72	<b>0.70</b>	0.75	<b>0.82</b>	0.83	<b>0.81</b>	0.84	0.73
e-rater	0.82	0.74	0.69	0.72	0.80	0.81	0.75	0.81	0.70
IntelliMetric	0.78	0.70	0.68	0.73	0.79	0.83	0.76	0.81	0.68
CRASE	0.76	0.72	0.69	0.73	0.76	0.78	0.78	0.80	0.68
LightSIDE	0.79	0.70	<b>0.63</b>	0.74	0.81	0.81	0.76	0.77	0.65
AutoScore	0.78	0.68	0.66	0.72	0.75	0.82	0.76	0.67	0.69
IEA	0.79	0.70	<b>0.65</b>	<b>0.65</b>	0.74	0.80	0.75	0.77	0.69
Bookette	0.70	0.68	0.63	0.69	0.76	0.80	0.64	0.74	0.60
Lexile	0.66	0.62	0.55	0.65	0.67	0.64	0.65	0.58	0.63

Primerjava state-of-the art sistemov<sup>1</sup> s sistemom SAGE+; uporaba kvadratne utežene Kappe.

1. M. D. Shermis and B. Hamner, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis," in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. Burstein, Eds. New York: Routledge, 2013, ch. 19, pp. 313–346.

# Hvala!

---

kaja.zupanc@fri.uni-lj.si