Text Similarity using Google Tri-grams

Aminul Islam, Evangelos Milios, and Vlado Keselj

DALHOUSIE UNIVERSITY
Inspiring Minds

Faculty of Computer Science
Dalhousie University, Canada

May 28, 2012, Canadian AI 2012, Toronto

Outline

Outline

Text Similarity

Text Similarity

P="An autograph is the signature of someone famous which is specially written for a fan to keep."

Text Similarity

P="An autograph is the signature of someone famous which is specially written for a fan to keep."

R="Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says."

Motivation

Motivation

- To develop an unsupervised method

Motivation

- To not use any lexical resources

Motivation

- To cover a large fragment of the language lexicon

Motivation

- To have language independence

Motivation

- To develop an unsupervised method
- To not use any lexical resources
- To cover a large fragment of the language lexicon
- To have language independence

Resource Used

Resource Used

Google Web 1T n-gram Corpus

Resource Used

Google Web 1T n-gram Corpus

Number of	Number	Size on disk (in KB)
Tokens	1,024,908,267,229	N/A
Sentences	95,119,665,584	N/A
Unigrams	13,588,391	185,569
Bigrams	314,843,401	5,213,440
Trigrams	977,069,902	19,978,540
4-grams	1,313,818,354	32,040,884
5-grams	1,176,470,663	33,678,504

Examples of Google n-gram Corpus

n=	n-grams	Frequencies
3	he was a	3,683,417
	hehe was a	52
	he was an	563,471
	he was am	121
	he was awesome	7,520
	he was awsome	548
4	he was a vegetarian	1,357
	he was a veritable	454
	he was a very	65,325
	he was a veteran	2,979
5	he was a very generous	276
	he was a very genuine	58
	he was a very gifted	177
	he was a very good	7,447

Outline

Text Similarity

P="An autograph is the signature of someone famous which is specially written for a fan to keep."

Text Similarity

P="An autograph is the signature of someone famous which is specially written for a fan to keep."

R="Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says."

Text Similarity

P="An autograph is the signature of someone famous which is specially written for a fan to keep."

R="Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says."

Step 1: (Preprocessing)

Text Similarity

P="An autograph is the signature of someone famous which is specially written for a fan to keep."

R="Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says."

Step 1: (Preprocessing)

Text Similarity

$P = \{\texttt{autograph}, \texttt{signature}, \texttt{famous}, \texttt{specially}, \texttt{written}, \texttt{fan}\}$

R="Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says."

Step 1: (Assign the length of P)

$m = |P| = 6$

Text Similarity

P = {autograph, signature, famous, specially, written, fan}

R = {signature, written, characteristic, end, document, wrote, document, agree}
$m = 6$

Step 1: (Assign the length of R)

$n = |R| = 8$

Text Similarity

P = {autograph, signature, famous, specially, written, fan}

R = {signature, written, characteristic, end, document, wrote, document, agree}
$m = 6$, $n = 8$

Step 1: (Find common words in P and R)

Text Similarity

P = {autograph, signature, famous, specially, written, fan}

R = {signature, written, characteristic, end, document, wrote, document, agree}
$m = 6$, $n = 8$

Step 1: (Find common words in P and R)

Text Similarity

P = {autograph, signature, famous, specially, written, fan}

R = {signature, written, characteristic, end, document, wrote, document, agree}
$m = 6$, $n = 8$

Step 1: (Find common words in P and R)
$\delta = 2$

Text Similarity

P = {autograph, ~~signature~~, famous, specially, ~~written~~, fan}

R = {~~signature~~, ~~written~~, characteristic, end, document, wrote, document, agree}
$m = 6$, $n = 8$, $\delta = 2$

Step 2: (Remove common words)

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
$m = 6$, $n = 8$, $\delta = 2$

Step 3: Construct a 4×6 'similarity matrix'

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
$m = 6$, $n = 8$, $\delta = 2$

Step 3: Construct a 4×6 'similarity matrix', M,

$$M = \begin{array}{c} \\ autograph \\ famous \\ specially \\ fan \end{array} \begin{array}{c} \begin{array}{cccccc} characteristic & end & document & wrote & document & agree \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2

Step 4:

$$M = \begin{array}{c} \\ autograph \\ famous \\ specially \\ fan \end{array} \begin{array}{c} \begin{array}{cccccc} characteristic & end & document & wrote & document & agree \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2

Step 4:

$\rho = \{0.282\}$

$$M = \begin{array}{c} \\ autograph \\ famous \\ specially \\ fan \end{array} \begin{array}{c} \begin{array}{cccccc} characteristic & end & document & wrote & document & agree \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2

Step 4:

$\rho = \{0.282\}$

$$M = \begin{array}{c} \\ \textit{autograph} \\ \textit{famous} \\ \textit{specially} \\ \textit{fan} \end{array} \begin{array}{c} \begin{array}{cccccc} \textit{characteristic} & \textit{end} & \textit{document} & \textit{wrote} & \textit{document} & \textit{agree} \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2

Step 4:

$\rho = \{0.282, 0.374\}$

$$M = \begin{array}{c} \\ autograph \\ famous \\ specially \\ fan \end{array} \begin{array}{c} \begin{array}{cccccc} characteristic & end & document & wrote & document & agree \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2

Step 4:

$\rho = \{0.282, 0.374\}$

$$M = \begin{array}{c} \\ \textit{autograph} \\ \textit{famous} \\ \textit{specially} \\ \textit{fan} \end{array} \begin{array}{c} \begin{array}{cccccc} \textit{characteristic} & \textit{end} & \textit{document} & \textit{wrote} & \textit{document} & \textit{agree} \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2

Step 4:

$\rho = \{0.282, 0.374, \frac{0.258+0.258}{2}\}$

$$M = \begin{array}{c} \\ autograph \\ famous \\ specially \\ fan \end{array} \begin{array}{c} \begin{array}{cccccc} characteristic & end & document & wrote & document & agree \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2

Step 4:

$\rho = \{0.282, 0.374, 0.258\}$

$$M = \begin{array}{c} \\ autograph \\ famous \\ specially \\ fan \end{array} \begin{array}{c} \begin{array}{cccccc} characteristic & end & document & wrote & document & agree \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2

Step 4:

$\rho = \{0.282, 0.374, 0.258, \frac{0.203+0.174}{2}\}$

$$M = \begin{array}{c} \\ autograph \\ famous \\ specially \\ fan \end{array} \begin{array}{c} \begin{array}{cccccc} characteristic & end & document & wrote & document & agree \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2

Step 4:

$\rho = \{0.282, 0.374, 0.258, 0.189\}$

$$M = \begin{array}{c} \\ autograph \\ famous \\ specially \\ fan \end{array} \begin{array}{c} \begin{array}{cccccc} characteristic & end & document & wrote & document & agree \end{array} \\ \begin{pmatrix} 0 & 0 & 0.259 & 0.282 & 0.259 & 0 \\ 0.257 & 0.055 & 0.051 & 0.374 & 0.051 & 0.001 \\ 0 & 0.168 & 0.258 & 0.137 & 0.258 & 0 \\ 0 & 0.012 & 0 & 0.203 & 0 & 0.174 \end{pmatrix} \end{array}$$

Text Similarity

P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2, ρ= {0.282, 0.374, 0.258, 0.189}

Step 5:

$$S(P,R) = \frac{(\delta + \sum_{i=1}^{|\rho|} \rho) \times (m+n)}{2mn}$$

Text Similarity

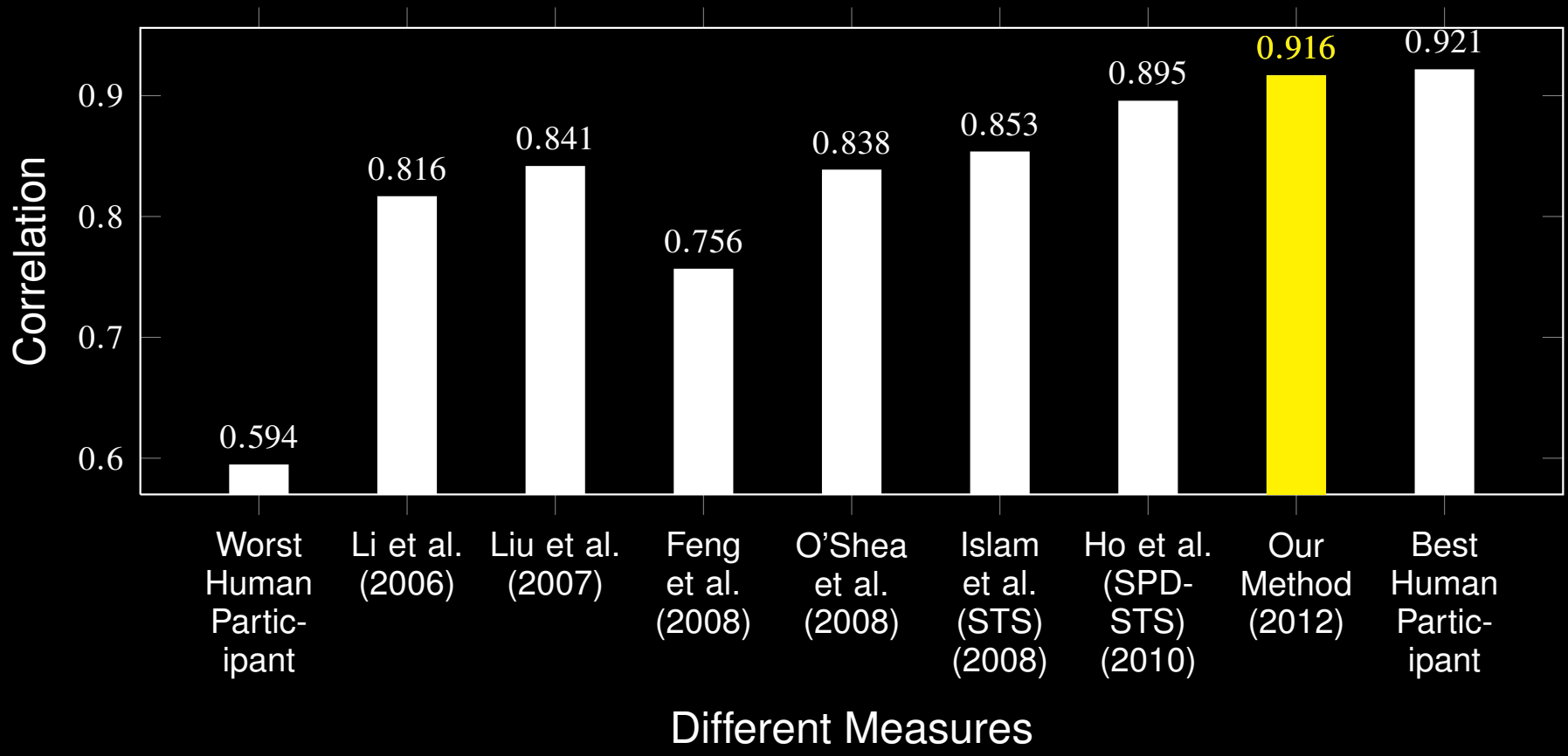
P = {autograph, famous, specially, fan}

R = {characteristic, end, document, wrote, document, agree}
m= 6, n= 8, δ= 2, ρ= {0.282, 0.374, 0.258, 0.189}

Step 5:

$$S(P,R) = \frac{(\delta + \sum_{i=1}^{|\rho|} \rho) \times (m+n)}{2mn}$$
$$= 0.447$$

Evaluation and Experimental Results



Outline

Conclusion

- Comparable to supervised methods

Conclusion

- Comparable to supervised methods
- The performance of our method is close to that of the best human participant

Conclusion

- Comparable to supervised methods
- The performance of our method is close to that of the best human participant
- To test our method with long documents and in other possible applications

Thanks