

Análise de Redes Sociais

Atribuição de Autoria

Prof. Luciano Antonio Digiampietri
EACH-USP

Atribuição de Autoria

Atribuição de autoria é a tarefa de se identificar o autor de um “item”.

Atribuição de Autoria

Atribuição de autoria é a tarefa de se identificar o autor de um “item”.

Tipicamente é tratada a autoria de um texto, mas o problema se aplica a códigos-fonte, pinturas, composições etc.

Atribuição de Autoria



***“May the force
be with you.”***

Gandalf

Atribuição de Autoria

Nesta aula será dado enfoque na atribuição de autoria em postagens em redes sociais.

Atribuição de Autoria

Nesta aula será dado enfoque na atribuição de autoria em postagens em redes sociais.

Focaremos no conteúdo dos textos, mas discutiremos também algumas informações adicionais da rede que podem auxiliar na tarefa.

Atribuição de Autoria

A atribuição de autoria pode ter diversas motivações:

Atribuição de Autoria

A atribuição de autoria pode ter diversas motivações:

Garantias de reconhecimento/*copyright*;

Atribuição de Autoria

A atribuição de autoria pode ter diversas motivações:

Garantias de reconhecimento/*copyright*;

Resolução de reivindicações;

Atribuição de Autoria

A atribuição de autoria pode ter diversas motivações:

- Garantias de reconhecimento/*copyright*;

- Resolução de reivindicações;

- Identificação de perfis banidos;

Atribuição de Autoria

A atribuição de autoria pode ter diversas motivações:

- Garantias de reconhecimento/*copyright*;

- Resolução de reivindicações;

- Identificação de perfis banidos;

- Resolução de crimes (*linguística forense*).

Atribuição de Autoria

Um dos casos mais famosos é a atribuição de autoria aos *The Federalist Papers*:

Atribuição de Autoria

Um dos casos mais famosos é a atribuição de autoria aos *The Federalist Papers*:

Coleção de 85 artigos e ensaios escritos sob o pseudônimo “Publius” (em 1787 e 1788) por Alexander Hamilton, James Madison e John Jay para promover a ratificação da Constituição dos Estados Unidos.

Atribuição de Autoria

Um dos casos mais famosos é a atribuição de autoria aos *The Federalist Papers*:

¹Mosteller e Wallace, (2012)

Atribuição de Autoria

Um dos casos mais famosos é a atribuição de autoria aos *The Federalist Papers*:

Em 1792 a lista de autores tornou-se pública;

¹Mosteller e Wallace, (2012)

Atribuição de Autoria

Um dos casos mais famosos é a atribuição de autoria aos *The Federalist Papers*:

Em 1792 a lista de autores tornou-se pública;
73 artigos tinham sua autoria “bem conhecida”;

¹Mosteller e Wallace, (2012)

Atribuição de Autoria

Um dos casos mais famosos é a atribuição de autoria aos *The Federalist Papers*:

Em 1792 a lista de autores tornou-se pública;
73 artigos tinham sua autoria “bem conhecida”;
Em 1944 foi realizada a atribuição atual de autoria
(confirmada por métodos computacionais em
1964)¹.

¹Mosteller e Wallace, (2012)

Atribuição de Autoria

Outras análises clássicas:

Atribuição de Autoria

Outras análises clássicas:

Romance Ulisses de James Joyce (escrito entre 1914 e 1921):

Atribuição de Autoria

Outras análises clássicas:

Romance Ulisses de James Joyce (escrito entre 1914 e 1921): **apresenta cinco estilos de escrita, todos diferentes do estilo do autor em sua primeira obra.**

Atribuição de Autoria

Outras análises clássicas:

Romance Ulisses de James Joyce (escrito entre 1914 e 1921): **apresenta cinco estilos de escrita, todos diferentes do estilo do autor em sua primeira obra.**

Peça *Double Falsehood* de Lewis Theobald:

Atribuição de Autoria

Outras análises clássicas:

Romance Ulisses de James Joyce (escrito entre 1914 e 1921): **apresenta cinco estilos de escrita, todos diferentes do estilo do autor em sua primeira obra.**

Peça *Double Falsehood* de Lewis Theobald: **possível “adaptação” de obra de John Fletcher e William Shakespeare.**

Atribuição de Autoria

Outras análises clássicas²:

A Doação de Constantino (*Donatio Constantini*):
decreto imperial romano do imperador
Constantino, o Grande, do século IV, transferindo
a autoridade sobre Roma e a parte ocidental do
Império Romano para o Papa:

²<https://en.wikipedia.org/wiki/Stylometry>

Atribuição de Autoria

Outras análises clássicas²:

A Doação de Constantino (*Donatio Constantini*):
decreto imperial romano do imperador
Constantino, o Grande, do século IV, transferindo
a autoridade sobre Roma e a parte ocidental do
Império Romano para o Papa: **potencialmente
falsificado.**

²<https://en.wikipedia.org/wiki/Stylometry>

Estilometria

Estilometria

Análise ‘estatística’ do estilo de um autor.

Estilometria

Análise 'estatística' do estilo de um autor.

Tipicamente aplicada ao estilo de escrita, mas pode ser utilizada para pinturas, códigos etc.

Estilometria

Análise ‘estatística’ do estilo de um autor.

Tipicamente aplicada ao estilo de escrita, mas pode ser utilizada para pinturas, códigos etc.

Comumente usada para a identificação do autor de textos anônimos ou de textos disputados.

Estilometria - Características

Estilometria - Características

Léxicas: baseadas nas palavras.

Estilometria - Características

Léxicas: baseadas nas palavras.

Riqueza do vocabulário;

Estilometria - Características

Léxicas: baseadas nas palavras.

Riqueza do vocabulário;

Frequência do uso das palavras;

Estilometria - Características

Léxicas: baseadas nas palavras.

Riqueza do vocabulário;

Frequência do uso das palavras;

Tamanho das palavras;

Estilometria - Características

Léxicas: baseadas nas palavras.

Riqueza do vocabulário;

Frequência do uso das palavras;

Tamanho das palavras;

Ordem das palavras;

Estilometria - Características

Léxicas: baseadas nas palavras.

Riqueza do vocabulário;

Frequência do uso das palavras;

Tamanho das palavras;

Ordem das palavras;

Erros ortográficos;

Estilometria - Características

Léxicas: baseadas nas palavras.

Riqueza do vocabulário;

Frequência do uso das palavras;

Tamanho das palavras;

Ordem das palavras;

Erros ortográficos;

Ordem das palavras*;

Estilometria - Características

Léxicas: baseadas nas palavras.

Riqueza do vocabulário;

Frequência do uso das palavras;

Tamanho das palavras;

Ordem das palavras;

Erros ortográficos;

Ordem das palavras*;

Tamanho das sentenças*.

Estilometria - Características

Sintáticas: baseadas na estrutura das frases e funções das palavras.

Estilometria - Características

Sintáticas: baseadas na estrutura das frases e funções das palavras.

Frequência do uso das funções (adjetivos, substantivos, artigos etc.);

Estilometria - Características

Sintáticas: baseadas na estrutura das frases e funções das palavras.

Frequência do uso das funções (adjetivos, substantivos, artigos etc.);

Construções frequentes (ordem das funções);

Estilometria - Características

Sintáticas: baseadas na estrutura das frases e funções das palavras.

Frequência do uso das funções (adjetivos, substantivos, artigos etc.);

Construções frequentes (ordem das funções);

Construções frequentes (uso de voz passiva, ativa);

Estilometria - Características

Sintáticas: baseadas na estrutura das frases e funções das palavras.

Frequência do uso das funções (adjetivos, substantivos, artigos etc.);

Construções frequentes (ordem das funções);

Construções frequentes (uso de voz passiva, ativa);

Erros sintáticos.

Estilometria - Características

Semânticas: significados das palavras (ou de trechos do texto).

Estilometria - Características

Semânticas: significados das palavras (ou de trechos do texto).

Significado das palavras utilizadas;

Estilometria - Características

Semânticas: significados das palavras (ou de trechos do texto).

Significado das palavras utilizadas;

Significado das sentenças;

Estilometria - Características

Semânticas: significados das palavras (ou de trechos do texto).

Significado das palavras utilizadas;

Significado das sentenças;

Conteúdo/assunto que está sendo discutido.

Outras Características

Outras: frequência de uso de caracteres,
pontuações, *emojis* etc.

Outras Características

Outras: frequência de uso de caracteres, pontuações, *emojis* etc.

Frequência (de conjuntos) de caracteres;

Outras Características

Outras: frequência de uso de caracteres, pontuações, *emojis* etc.

Frequência (de conjuntos) de caracteres;

Frequência e local das pontuações;

Outras Características

Outras: frequência de uso de caracteres, pontuações, *emojis* etc.

Frequência (de conjuntos) de caracteres;

Frequência e local das pontuações;

Frequência do uso de cada emoji;

Outras Características

Outras: frequência de uso de caracteres, pontuações, *emojis* etc.

Frequência (de conjuntos) de caracteres;

Frequência e local das pontuações;

Frequência do uso de cada emoji;

Grupos em que as mensagens são publicadas;

Outras Características

Outras: frequência de uso de caracteres, pontuações, *emojis* etc.

Frequência (de conjuntos) de caracteres;

Frequência e local das pontuações;

Frequência do uso de cada emoji;

Grupos em que as mensagens são publicadas;

Horários e intervalos entre publicações.

Linguística Forense

³http://www.forensiclinguistics.net/cfl_fl.html

Linguística Forense

Aplicação do conhecimento linguístico no contexto da lei, investigação criminal, julgamento e processos judiciais.

³http://www.forensiclinguistics.net/cfl_fl.html

Linguística Forense

Aplicação do conhecimento linguístico no contexto da lei, investigação criminal, julgamento e processos judiciais.

Entendimento da lei (escrita); ³

³http://www.forensiclinguistics.net/cfl_fl.html

Linguística Forense

Aplicação do conhecimento linguístico no contexto da lei, investigação criminal, julgamento e processos judiciais.

Entendimento da lei (escrita); ³

Compreensão do uso da linguagem em processos judiciais;

³http://www.forensiclinguistics.net/cfl_fl.html

Linguística Forense

Aplicação do conhecimento linguístico no contexto da lei, investigação criminal, julgamento e processos judiciais.

Entendimento da lei (escrita); ³

Compreensão do uso da linguagem em processos judiciais;

Fornecimento de provas linguísticas.

³http://www.forensiclinguistics.net/cfl_fl.html

Linguística Forense

As evidências da linguística forense têm sido mais utilizadas para inocentar uma pessoa do que para ‘provar’ que é culpada.

Linguística Forense

As evidências da linguística forense têm sido mais utilizadas para inocentar uma pessoa do que para ‘provar’ que é culpada.

Mas há casos em que as evidências linguísticas foram usadas para (ajudar a) incriminar suspeitos.

Linguística Forense

As evidências da linguística forense têm sido mais utilizadas para inocentar uma pessoa do que para ‘provar’ que é culpada.

Mas há casos em que as evidências linguísticas foram usadas para (ajudar a) incriminar suspeitos.

Linguística Forense

As evidências da linguística forense têm sido mais utilizadas para inocentar uma pessoa do que para ‘provar’ que é culpada.

Mas há casos em que as evidências linguísticas foram usadas para (ajudar a) incriminar suspeitos. Um caso famoso é do *Unabomber*.

Atribuição de Autoria

Soluções computacionais tipicamente utilizam três abordagens:

Atribuição de Autoria

Soluções computacionais tipicamente utilizam três abordagens:

Detecção de Anomalias (*Outliers*):

Atribuição de Autoria

Soluções computacionais tipicamente utilizam três abordagens:

Detecção de Anomalias (*Outliers*): cria-se um modelo para um autor e verifica se o texto é ‘compatível’ com este modelo;

Atribuição de Autoria

Soluções computacionais tipicamente utilizam três abordagens:

Detecção de Anomalias (*Outliers*): cria-se um modelo para um autor e verifica se o texto é ‘compatível’ com este modelo;

Agrupamento:

Atribuição de Autoria

Soluções computacionais tipicamente utilizam três abordagens:

Detecção de Anomalias (*Outliers*): cria-se um modelo para um autor e verifica se o texto é ‘compatível’ com este modelo;

Agrupamento: agrupa-se os textos de acordo com o estilo (mesmo se os autores de todos são desconhecidos);

Atribuição de Autoria

Soluções computacionais tipicamente utilizam três abordagens:

Classificação:

Atribuição de Autoria

Soluções computacionais tipicamente utilizam três abordagens:

Classificação: a partir de um conjunto de textos de diferentes autores, um modelo de classificação é construído para classificar novos textos.

Atribuição de Autoria

Para estas três abordagens, diferentes características de cada texto são extraídas (Mineração de Textos / Processamento de Língua Natural).

Atribuição de Autoria

Considerando tanto postagens em redes sociais como textos mais longos, algumas soluções têm atingido excelentes resultados:

Atribuição de Autoria

Considerando tanto postagens em redes sociais como textos mais longos, algumas soluções têm atingido excelentes resultados:

Classificação: usando algoritmos como *Random Forest* e SVM;

Atribuição de Autoria

Considerando tanto postagens em redes sociais como textos mais longos, algumas soluções têm atingido excelentes resultados:

Classificação: usando algoritmos como *Random Forest* e SVM;

Usando: n-gramas de caracteres com n de 4 a 6;

Atribuição de Autoria

Considerando tanto postagens em redes sociais como textos mais longos, algumas soluções têm atingido excelentes resultados:

Classificação: usando algoritmos como *Random Forest* e SVM;

Usando: n-gramas de caracteres com n de 4 a 6;

N-gramas de palavras costumam apresentar resultados satisfatórios para textos ‘maiores’ (com n de 2 a 3).

Referências

- 1 A. Rocha et al. (2017), Authorship Attribution for Social Media Forensics, in IEEE Transactions on Information Forensics and Security, vol. 12, no. 1, pp. 5-33, doi: 10.1109/TIFS.2016.2603960.
- 2 Mosteller, Frederick; Wallace, David L. (2012). Applied Bayesian and Classical Inference: The Case of The Federalist Papers. Springer Science & Business Media. ISBN 978-1-4612-5256-6.
<https://doi.org/10.1002/asi.21001>
- 3 Jack Grieve, (2007). Quantitative Authorship Attribution: An Evaluation of Techniques, Literary and Linguistic Computing, Volume 22, Issue 3, September 2007, Pages 251–270, <https://doi.org/10.1093/llc/fqm020>

Análise de Redes Sociais

Atribuição de Autoria

Prof. Luciano Antonio Digiampietri
EACH-USP