

# MAL-FITT

## MyAnimeList Forum Interpreter Through Text

*ENIAC 2016 - Encontro Nacional de Inteligência Artificial e Computacional*

Gianlucca Lodron Zuin

Luiz Felipe Gonçalves Magalhães

Túlio Correa Loures



UNIVERSIDADE FEDERAL  
DE MINAS GERAIS



# Introdução

- Em uma discussão, temos frequentemente grande quantidade de temas e tópicos discutidos:
  - Para alguém externo, isso pode significar em uma barreira para o entendimento da ideia geral que está sendo tratada.
  - Seria interessante poder apresentar para ele a ideia central, aquilo que resume a discussão.
- Esse é um ponto especialmente frequente em fóruns de discussões diversos presentes na Web: Grande quantidade de postagens em um mesmo tópico, muitas vezes com conteúdos extremamente divergentes.



# Objetivo

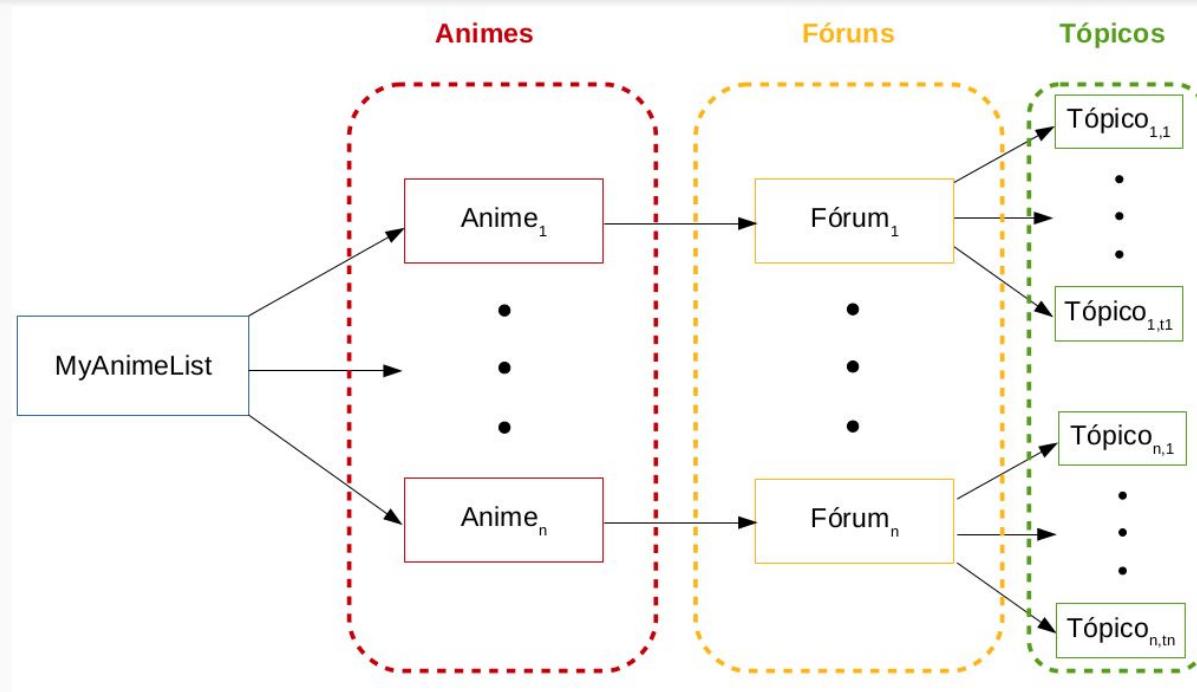
- Elaborar um método que retorne ao usuário o comentário central, que melhor resume a ideia de determinado tópico de um fórum de discussão.
- Evitar o uso de indicadores de “importância”, como curtidas (facebook) e upvotes (reddit):
  - Retratam postagens mais populares, mas não necessariamente mais representativas do tópico.

# Fórum WEB Abordado

- Portal [MyAnimeList.net](#) (MAL)
  - Página dedicada a permitir que usuários montem listas dos animes e mangás os quais pretendem assistir, estão assistindo ou já assistiram.
  - Mais de 3.000.000 Usuários, 48.000 séries registradas, e 655.000 tópicos de discussão.
  - Cada Anime ou Mangá possui seu próprio fórum de discussão, com tópicos discutindo assuntos diversos.

**MyAnimeList**

# Estrutura do MAL



# Tratamento dos Dados

- Coleta do texto não estruturado de um tópico.
- Remoção de caracteres não-Alfanuméricos.
- Normatização para letra minúscula.
- Remoção de stopwords.

# Método de Sentimento Proposto

- Associa a cada entidade foco de uma sentença a sua polaridade no contexto do comentário.
  - Foco: substantivos e entidades nomeadas.
  - Polaridade: parcela positiva, neutra e negativa retornada pela análise do VADER.
- Busca a postagem que melhor se assemelha à tendência do tópico (ie.: aborda os temas mais comentados e com uma polaridade similar aos demais comentários).
- Usa a saída do VADER como coordenadas no espaço tridimensional e calcula a distância de cosseno entre entidades.

# Método de Sentimento Proposto

O quanto interessante um termo é para a summarização da discussão pode ser encontrado pelo produto da similaridade entre sua polaridade e média no tópico pela frequência que o termo é utilizado.

$$dist_{x=termo} = \cos(\theta) = \frac{\sum pol_x pol_{\bar{x}}}{\sqrt{\sum pol_x^2} \sqrt{\sum pol_{\bar{x}}^2}} \quad (3)$$

$$fitness(x = termo) = sim_x * frequencia_x = (1 - dist_x) * frequencia_x \quad (4)$$

A aptidão da postagem, por sua vez, pode ser calculada pela média da aptidão de cada uma de suas sentenças (que é calculada pela soma das aptidões dos termos focos)

$$fitness(p = postagem) = \frac{\sum_{s \in S} fitness(s)}{|S|} = \frac{\sum_{s \in S} \sum_{f \in F} fitness(f)}{|S|} \quad (5)$$

# Método TF-IDF

- Considerar cada postagem do tópico como um documento.
- Criação de um documento adicional:
  - Concatenação de todas as postagens do tópico.
  - Visa representar a discussão como um todo.
- Formar de uma coleção (Tópico) com diversos documentos (Postagens)
- Usar TF-IDF para comparar a concatenação com os demais. A postagem com a maior similaridade sumariza a discussão.

# Método TF-IDF

Três métricas utilizadas para encontrar a postagem mais relevante:

- Maior similaridade com a concatenação dos documentos (TFIDF\_COMM)
  - Inserido ao fim da coleção um documento que é a concatenação de todos os anteriores.
  - A hipótese é que postagens próximas a ele melhor summarizam o tópico.

# Método TF-IDF

Três métricas utilizadas para encontrar a postagem mais relevante:

- Maior similaridade média com todos os outros documentos (TFIDF\_AVGC) Equação 1
- Maior similaridade mínima com todos os outros documentos (TFIDF\_MAXC) Equação 2
  - Calculam a similaridade par-a-par entre os documentos.

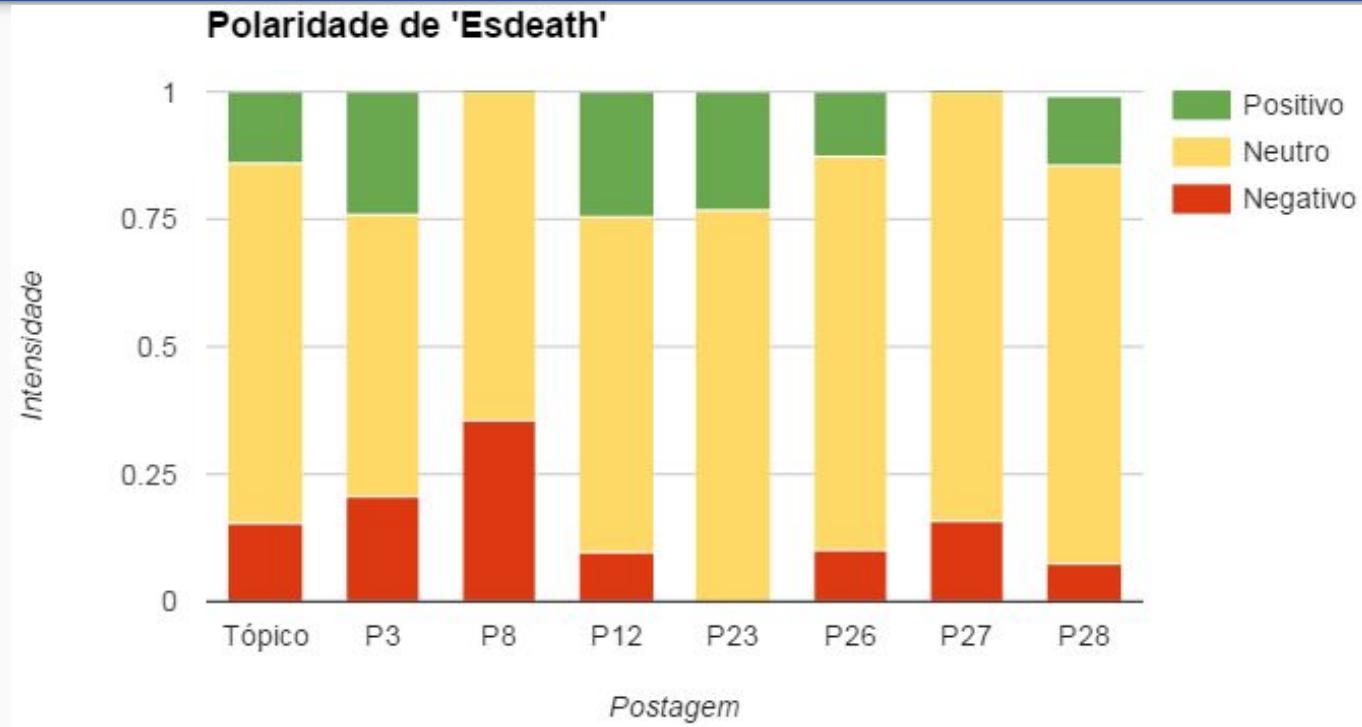
$$f_{avgc}(T = \text{topico}) = \max_{i \in T} \left( \frac{1}{|T|} * \sum_{j \in T, j \neq i} \text{similaridade}(i, j) \right) \quad (1)$$

$$f_{maxc}(T = \text{topico}) = \operatorname{argmax}_{i \in T} (\min_{j \in T, j \neq i} (\text{similaridade}(i, j))) \quad (2)$$

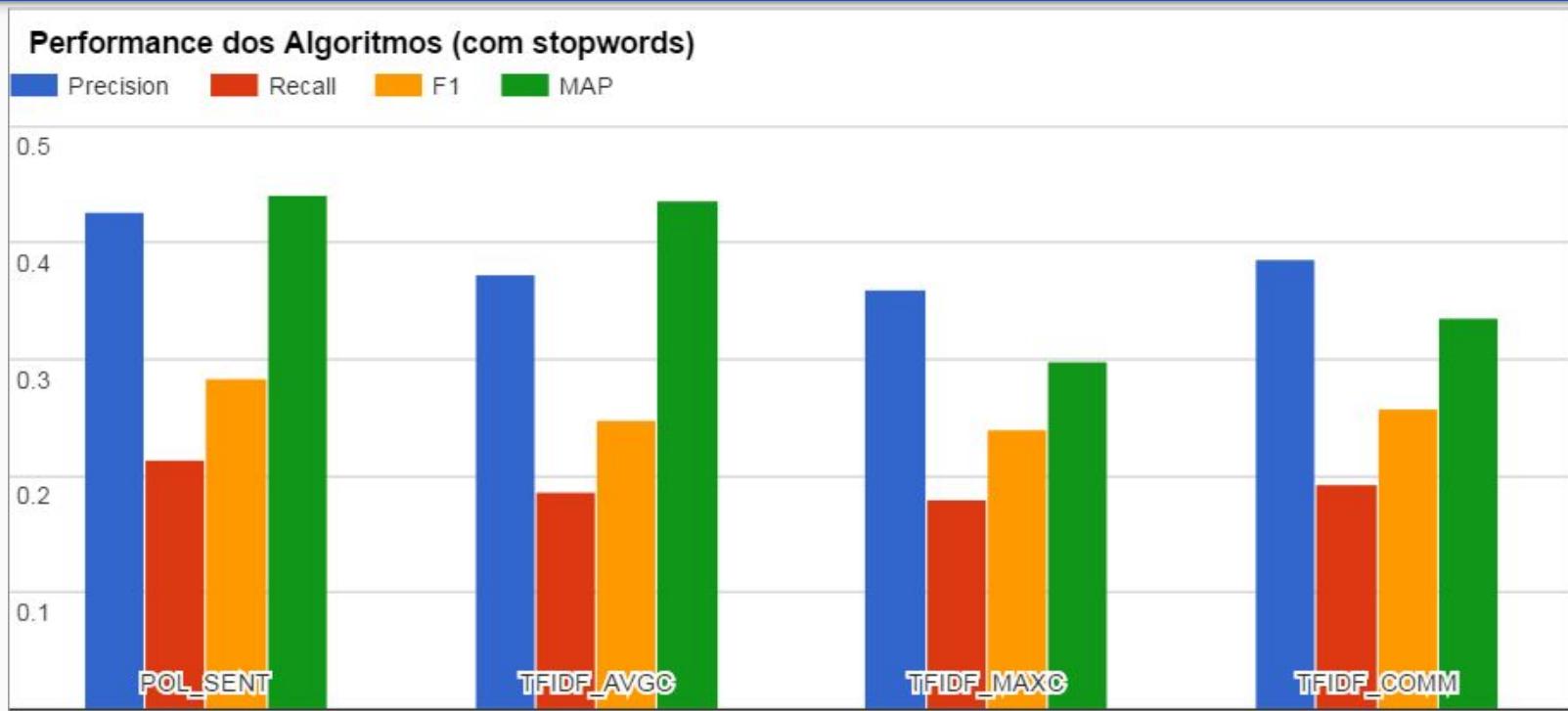
# Experimentos

- 15 tópicos de diferentes animes foram avaliados.
- Temas diversos (discussão sobre a trama, personagens, opinião, etc).
- Para cada tópico, selecionadas as 10 postagens mais relevantes:
  - Votação entre 3 avaliadores.
- Comparar a saída dos algoritmos (5 melhores postagens) com a humana.

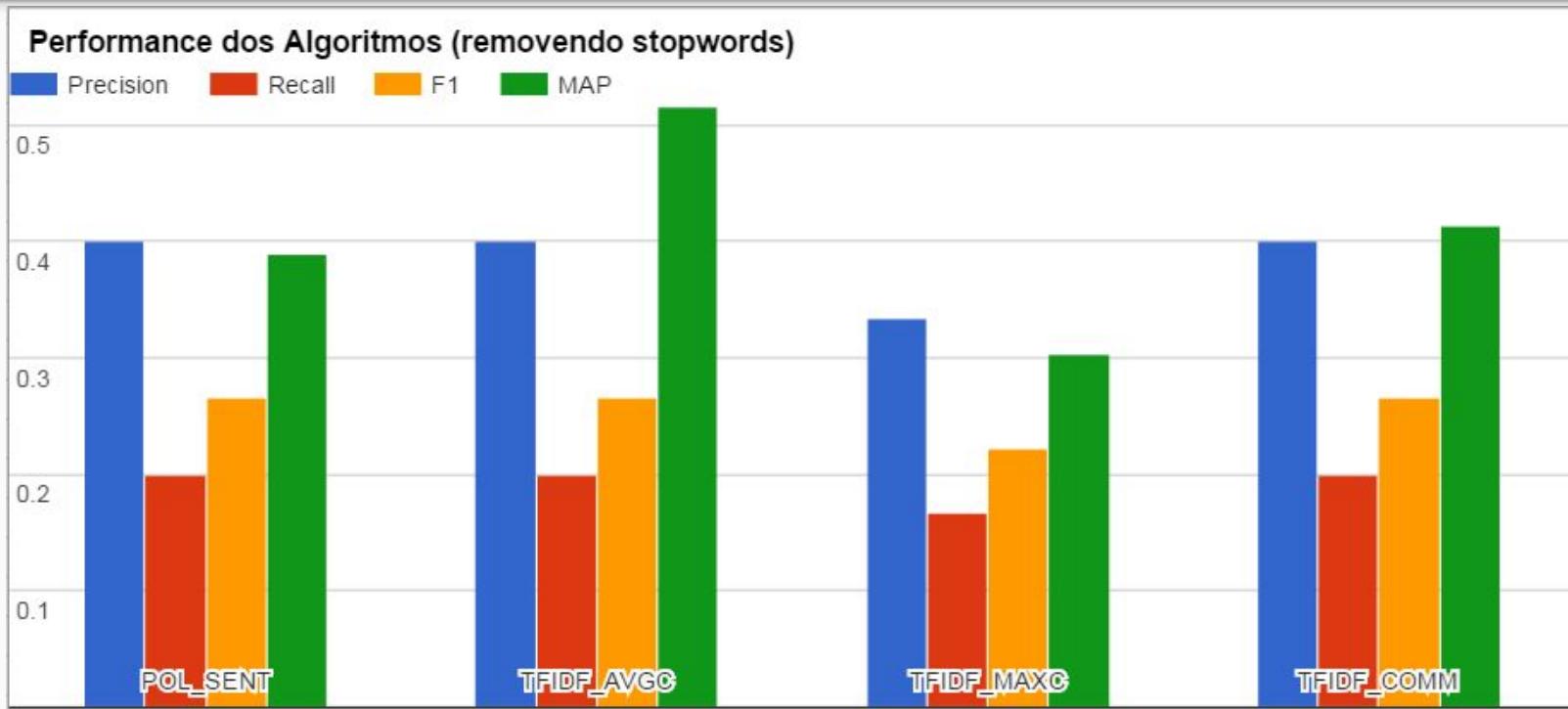
# Experimentos



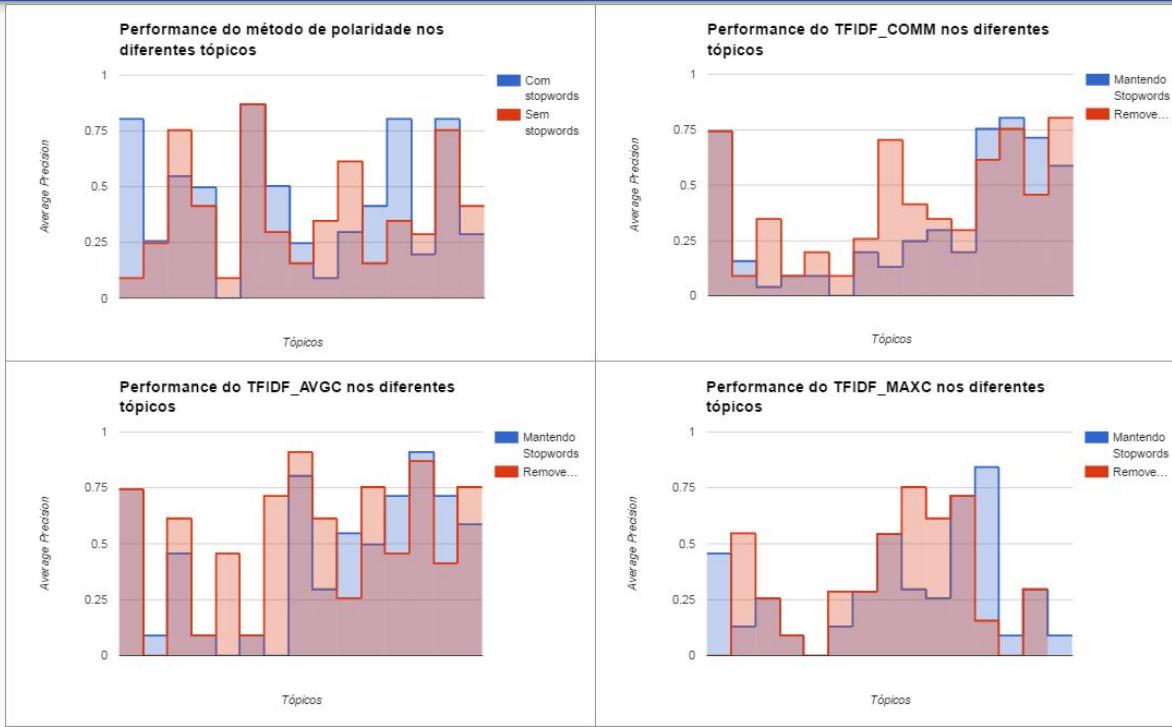
# Experimentos



# Experimentos



# Experimentos



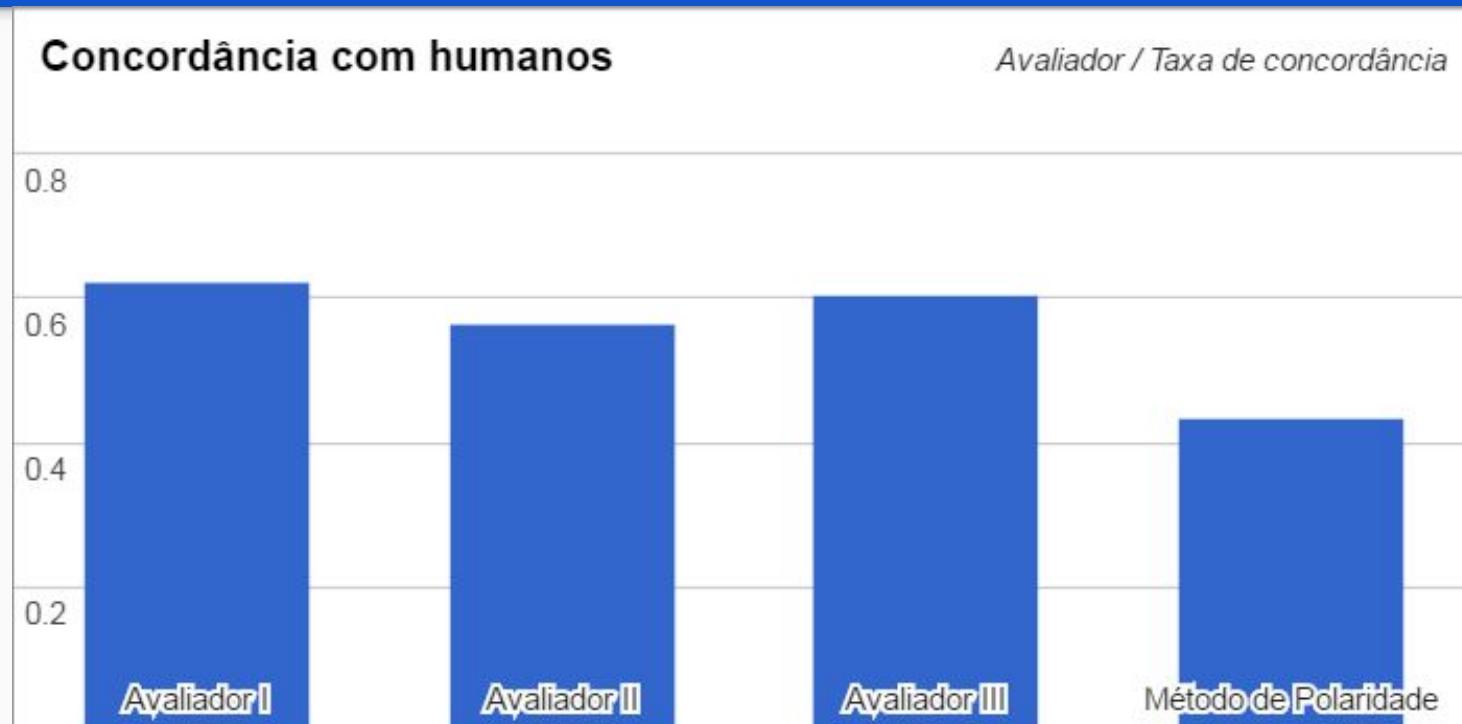
# Experimentos

- A performance ruim do TF-IDF e do método de sentimento em alguns tópicos pode ser justificada pelo uso de palavras similares. Os métodos não possuem a capacidade de abstrair o sentido semântico, palavras como '*graphics*' e '*visuals*' acabam sendo consideradas diferentes.
- Utilizada a *WordNet* para tentar colapsar termos com significados similares.
- Resultados pioraram, alguns termos foram colapsados erroneamente pois, dentre seus possíveis significados, um deles é comum. Além disso, *WordNet* mede o parentesco de palavras, e não necessariamente similaridade. '*Love*' está mais próximo de '*hate*' do que de '*passion*'.

# Conclusão

- Humanos, na média, concordam em cerca de 60%
- Melhores resultados: método de Polaridade
  - Polaridade concorda com humanos 43% das vezes
  - 73% do valor entre humanos
- Dentre as representações de TF-IDF, o melhor foi obtido com o uso da maior similaridade média

# Conclusão



# Conclusão

- Em casos onde muitos termos de significados similares eram utilizados, os métodos propostos não obtiveram bons resultados (tópico 2: ‘cgi’, ‘graphics’, ‘visuals’, etc).
  - Aplicação básica de WordNet não foi efetiva
- Em contrapartida, nos tópicos mais simples ele se mostrou extremamente efetivo
  - Tópico 13: 0.87 AP no TFIDF\_AVGC removendo stopwords



# Conclusão

- Possíveis estudos futuros:
  - Melhoria ao se considerar as citações (*quotes*) de comentários
  - Processo avaliativo mais abrangente
  - Considerar métricas de avaliação mais informativas
  - Considerar outras características relevantes para representação
    - Similaridade entre definição (*gloss*) das palavras
    - Confiabilidade do usuário na base
    - Utilizar outros dados disponíveis no MAL



# MAL-FITT

## MyAnimeList Forum Interpreter Through Text

*ENIAC 2016 - Encontro Nacional de Inteligência Artificial e Computacional*

Gianlucca Lodron Zuin

Luiz Felipe Gonçalves Magalhães

Túlio Correa Loures

