

PARTICIÓN DATASET PAN 2022

DATASET PAN 2022

- Para descargar el dataset se debe solicitar acceso al documento vía zenodo.
 - <https://pan.webis.de/>
 - https://zenodo.org/record/6337151#.YmjjW_PMleZ

ESTRUCTURA

- El conjunto de datos de entrenamiento viene con dos archivos JSON delimitados por saltos de línea codificados en UTF8. El primer archivo `pairs.jsonl` contiene pares de textos (cada par tiene una ID única) y sus etiquetas de tipo de discurso:
 - 1. {"id": "6cced668-6e51-5212-873c-717f2bc91ce6", "discourse_type": ["essay", "email"], "pair": ["Text 1...", "Text 2..."]}
 - 2. {"id": "ae9297e9-2ae5-5e3f-a2ab-ef7c322f2647", "discourse_type": ["email", "text_message"], "pair": ["Text 3...", "Text 4..."]}
 - 3. ...
- El segundo archivo, `Truth.jsonl`, contiene la verdad fundamental para todos los pares. La verdad básica se compone de una bandera booleana que indica si los textos en un par son del mismo autor y las identificaciones del autor:
 - 1. {"id": "6cced668-6e51-5212-873c-717f2bc91ce6", "same": true, "authors": ["1446633", "1446633"]}
 - 2. {"id": "ae9297e9-2ae5-5e3f-a2ab-ef7c322f2647", "same": false, "authors": ["1535385", "1998978"]}
 - 3. ...

INFORMACIÓN DEL DATASET

- Textos totales: 1,046
- Autores: 56
- Problemas: 12,101
 - Originalmente eran 12,264 pero se descartaron los que tenían las mismas parejas de textos.
- Tipos de discurso: 4
 - ['email', 'text_message', 'essay', 'memo']

CANTIDAD DE TEXTOS POR AUTOR

Se muestra el id del autor seguido de la cantidad de textos

- en_112: 39
- en_36: 38
- en_110: 34
- en_2: 29
- en_13: 28
- en_103: 26
- en_97: 26
- en_37: 25
- en_34: 23
- en_72: 22
- en_105: 22
- en_60: 22
- en_55: 22
- en_102: 21
- en_5: 21
- en_77: 21
- en_11: 20
- en_53: 20
- en_56: 19
- en_62: 19
- en_59: 19
- en_51: 18
- en_63: 18
- en_19: 18
- en_21: 18
- en_111: 18
- en_100: 18
- en_75: 18
- en_4: 18
- en_54: 17
- en_66: 17
- en_35: 17
- en_98: 16
- en_52: 16
- en_99: 16
- en_101: 16
- en_61: 16
- en_107: 16
- en_20: 16
- en_58: 16
- en_104: 15
- en_113: 15
- en_96: 15
- en_22: 15
- en_3: 14
- en_18: 14
- en_109: 13
- en_73: 13
- en_78: 13
- en_108: 13
- en_74: 12
- en_12: 11
- en_67: 11
- en_57: 11
- en_76: 11
- en_114: 11

REQUERIMIENTO

- Generar particiones del dataset para entrenamiento, validación y prueba.
- Consideraciones :
 - Agrupar en cada conjunto las listas de problemas que pertenecen a un grupo determinado de autores, de tal manera que cada conjunto sea ajeno en autores.
 - Asegurar que cada conjunto tenga la misma cantidad de problemas con resultado positivos y negativos.

PROBLEMÁTICA

- Al ejecutar la implementación realizada para el dataset del PAN 2020, no se pudieron hacer las particiones por autor ajeno, debido a que todos los autores están relacionados entre ellos.
- Esto es debido a la diferencia tan grande entre problemas y número de autores (12101 vs 56).

SOLUCIÓN #1

- Remover la validación de autor ajeno.
- Filtrar los textos con longitud menor a 500 tokens.
- Generar las particiones.

PARTICIONES SOLUCIÓN #1

- Entrenamiento
 - Total de problemas: 2,857 de los cuales 1,291 son positivos.
- Validación
 - Total de problemas: 354 de los cuales 175 son positivos.
- Prueba
 - Total de problemas: 356 de los cuales 165 son positivos.

VERIFICACIÓN SOLUCIÓN #1

- Intersecciones entre particiones:
 - Entrenamiento y validación: 180 textos y 52 autores.
 - Entrenamiento y prueba: 176 textos y 51 autores.
 - Prueba y validación: 67 textos y 51 autores.
- Total de problemas de las particiones: 3,567
 - Se eliminaron 8,534 problemas

SOLUCIÓN #2

- Remover el filtrado de los textos con longitud menor a 500 tokens.
- Generar grupos de autores ordenados con base al número de textos asociados a ellos.
 - Se ordenaron de esta forma para asignar los autores con mayor número de textos a la partición de entrenamiento.
- Remover todos los problemas en los que intervengan autores de diferente grupo.
- Generar las particiones.

PARTICIONES SOLUCIÓN #2

- Entrenamiento
 - Total de problemas: 5889 de los cuales 5469 son positivos.
- Validación
 - Total de problemas: 287 de los cuales 274 son positivos.
- Prueba
 - Total de problemas: 411 de los cuales 389 son positivos.

VERIFICACIÓN SOLUCIÓN #2

- Intersecciones entre particiones: no hubo coincidencias.
- Total de problemas de las particiones: 6587
 - Se eliminaron 5514.

SOLUCIÓN #3

- Remover el filtrado de los textos con longitud menor a 500 tokens.
- Generar grupos de autores ordenados con base al número de textos asociados a ellos.
 - Se ordenaron de esta forma para asignar los autores con mayor número de textos a la partición de entrenamiento.
- Remover todos los problemas en los que intervengan autores de diferente grupo.
- Generar las particiones.
- Crear nuevos problemas para cada partición.

- Para generar los nuevos problemas se realizó lo siguiente:
 - Se obtuvo el patrón de tipos de discurso que siguen los problemas del dataset.
 - Se utilizó un modelo generativo (clasificador bayesiano) para obtener las parejas de tipos de discurso, este clasificador se entrenó utilizando solo los tipos de discurso de cada problema del dataset completo.
 - Pattern = {'email':'text_message', 'memo':'email', 'essay':'email'}
 - Utilizando los autores de cada partición, el patrón de tipos de discurso y sus textos se generaron todas las combinaciones posibles de problemas positivos, y para el caso de los negativos se generaron todos los necesarios de tal manera que quedaran balanceados positivos y negativos.

PARTICIONES SOLUCIÓN #3

- Entrenamiento
 - Total de problemas: 15,732 de los cuales 7866 son positivos.
 - Se generaron 9,843 problemas nuevos de los cuales 2,397 son positivos.
- Validación
 - Total de problemas: 754 de los cuales 377 son positivos.
 - Se generaron 467 problemas nuevos de los cuales 103 son positivos.
- Prueba
 - Total de problemas: 1070 de los cuales 535 son positivos.
 - Se generaron 659 problemas nuevos de los cuales 146 son positivos.

VERIFICACIÓN SOLUCIÓN #3

- Intersecciones entre particiones: no hubo coincidencias.
- Total de problemas de las particiones: 17,556
 - Se eliminaron 5514 problemas del dataset original y se generaron 10,969 nuevos

EVALUACIÓN

SE EJECUTARON LOS
BASELINE CON
CADA UNO DE LOS
DATASET Y SE
REGISTRARON LAS
MÉTRICAS DE CADA
UNO

RESULTADOS

SOLUCIÓN #1								
Baseline	Train	Test	F1	AUC	Brier	c@1	f_05_u	overall
CosineSimilarity	2857	356	0.64	0.61	0.764	0.563	0.554	0.626
SVC	2857	356	0.129	0.443	0.469	0.469	0.188	0.34
TextCompression	2857	356	0.113	0.525	0.559	0.235	0.235	0.398

RESULTADOS

SOLUCIÓN #2

Baseline	Train	Test	F1	AUC	Brier	c@1	f_05_u	overall
CosineSimilarity	5889	411	0.97	0.427	0.806	0.941	0.954	0.82
SVC	5889	411	0.972	0.5	0.946	0.946	0.957	0.864
TextCompression	5889	411	0.972	0.482	0.946	0.946	0.957	0.861

RESULTADOS

SOLUCIÓN #3

Baseline	Train	Test	F1	AUC	Brier	c@1	f_05_u	overall
CosineSimilarity	15732	1070	0.665	0.442	0.691	0.498	0.554	0.57
SVC	15732	1070	0.696	0.575	0.575	0.575	0.594	0.603
TextCompression	15732	1070	0.667	0.474	0.564	0.5	0.556	0.552

CONCLUSIÓN

- Se eligió trabajar con la solución #3, debido a que cumple con las características apropiadas para entrenar un modelo, es decir, cuenta con registros de entrenamiento mayores a 10 mil, tiene el mismo número de problemas positivos como negativos en cada partición y no comparten autores
- Cabe mencionar que a pesar de que en la solución #2 se obtuvieron mejores resultados en las métricas de evaluación no se eligió debido a que puede deberse a que tanto en la partición de entrenamiento como en la de prueba, casi todos los registros son positivos.