

**Dataset and task description of the
information extraction tasks used for
evaluation in SIGIR 2020 submission id: 29**

The Medical Article Record Groundtruth Dataset

Dataset overview:

No. of document types: 9
No. of documents: 1553
Dataset properties: Scanned articles of biomedical journals from the NLM
Publicly available? Yes
Download link: [link](#)

Information extraction task overview:

No. of distinct named entities: 4
List of named entities: See below

| Index | Named entity | Description |
|-------|-----------------------|--|
| 1 | <i>"Title"</i> | Name of the article |
| 2 | <i>"Contribution"</i> | The main contribution of the scientific article (as claimed by the authors) |
| 3 | <i>"Authors"</i> | Name of the researchers who authored the article |
| 4 | <i>"Affiliations"</i> | The institutions each author belonged to during the publication of the article |

The Tobacco Litigation Dataset

Dataset overview:

No. of document types: 10
No. of documents: 3482
Dataset properties: Scanned documents from publicly available records of a lawsuit
Publicly available? Yes
Download link: [link](#)

Information extraction task overview:

No. of distinct named entities: 34
List of named entities: See below

| Index | Named entity | Description |
|-------|------------------------------------|--|
| 1 | <i>“Sender”</i> | The sender of an email/letter/memo |
| 2 | <i>“Receiver”</i> | The receiver of an email/letter/memo |
| 3 | <i>“Timestamp”</i> | Date of communication via an email/letter/memo |
| 4 | <i>“Receiver’s address”</i> | The listed address of the receiver of an email/letter |
| 5 | <i>“Sender’s affiliation”</i> | Position and affiliation of the sender of an email/letter |
| 6 | <i>“Subject of communication”</i> | The subject of an email/letter/memo |
| 7 | <i>“Reported tobacco product”</i> | Name of the tobacco product mentioned in a news/scientific report |
| 8 | <i>“Affiliated company’s name”</i> | Name of a tobacco product manufacturer mentioned in a news/scientific report |
| 9 | <i>“Action items”</i> | Todo’s as mentioned in an email/letter/memo |
| 10 | <i>“Action deadline”</i> | Deadlines of the todo list |
| 11 | <i>“News headline”</i> | Title of a news article/scientific report |
| 12 | <i>“Publication date”</i> | Date of publication of a news article/scientific report |

| Index | Named entity | Description |
|-------|---|---|
| 13 | <i>“Reported side effects”</i> | Harmful effects of tobacco usage as mentioned in a news article/scientific article |
| 14 | <i>“Age group of the affected population”</i> | The majority age group of the affected population |
| 15 | <i>“Potential lawsuit details”</i> | Plaintiffs of a potential lawsuit as mentioned in a news article |
| 16 | <i>“Plaintiff’s reported location”</i> | City/location of the plaintiffs of a potential lawsuit as mentioned in a news article |
| 17 | <i>“Recommended regulatory action”</i> | Potential prohibitory actions to be introduced as mentioned in a news article |
| 18 | <i>“Author”</i> | Contributors of a news article/scientific report |
| 19 | <i>“Authors’ affiliations”</i> | Affiliations of the contributors |
| 20 | <i>“Sources cited”</i> | References cited by a scientific report |
| 21 | <i>“Key findings”</i> | Main discoveries of the report |
| 22 | <i>“Date of publication”</i> | The date when a news article/scientific report was published |
| 23 | <i>“Job applicant”</i> | Name of the applicant applying for a job |
| 24 | <i>“Age”</i> | Age of the applicant as mentioned in the resume |
| 25 | <i>“Gender”</i> | Gender of the job applicant as mentioned in the resume |
| 26 | <i>“Position applied”</i> | Title of the applied job |
| 27 | <i>“Employment history”</i> | Past work experience of the applicant |
| 28 | <i>“Highest qualification”</i> | Highest educational qualification of the job applicant |
| 29 | <i>“Honors”</i> | Awards received by the applicant |
| 30 | <i>“Membership of societies”</i> | List of scientific societies the applicant is a member of |
| 31 | <i>“Publications”</i> | List of publications listed by the applicant in their resume |
| 32 | <i>“Report title”</i> | Title of a submitted scientific report |
| 33 | <i>“Report number”</i> | Serial no. of the report |

| Index | Named entity | Description |
|-------|--------------------------|--|
| 34 | <i>“Submission date”</i> | The date when the report was submitted |

The Brains Dataset

Dataset overview:

No. of document types: 36
No. of documents: ~ 1.4M
Dataset properties: Scanned documents for keeping vital patient information in ER
Publicly available? Will be made publicly available
Download link: N/A

Information extraction task overview:

No. of distinct named entities: 10
List of named entities: See below

| Index | Named entity | Description |
|-------|-----------------------------|--|
| 1 | <i>Patient's name</i> | Name of the patient under medical care |
| 2 | <i>Age</i> | Patient's age when admitted |
| 3 | <i>Gender</i> | Patient's gender |
| 4 | <i>Code</i> | Resuscitation status of the patient |
| 5 | <i>Admit date</i> | The day when the patient was first admitted to the ER |
| 6 | <i>Room number</i> | Room number where the patient is now in the hospital |
| 7 | <i>Diagnosis</i> | Latest diagnosis of the patient made by the medical doctor responsible for the patient |
| 8 | <i>Medical history</i> | Past medical records of the patient |
| 9 | <i>Dietary restrictions</i> | Known food allergies |
| 10 | <i>Consulting physician</i> | Name of the medical doctor responsible for the patient |

The NIST Special Dataset

Dataset overview:

| | |
|------------------------|---|
| No. of document types: | 20 |
| No. of documents: | 5595 |
| Dataset properties: | Tax documents from the IRS 1040 package of 1988 |
| Publicly available? | Yes |
| Download link: | link |

Information extraction task overview:

| | |
|---------------------------------|-----------|
| No. of distinct named entities: | 1379 |
| List of named entities: | See below |

Each of the form fields in this dataset corresponds to a named entity in the information extraction task defined on this dataset. A complete list of these named entities can be found at https://s3.amazonaws.com/nist-srd/SD6/SD06_users_guide.pdf