Основано на:

https://nlpub.github.io/russe-wsi-kit/

Нужно выкачать репозиторий, далее, используя данные из трейн сета \main\active-dict\train.csv , которое содержит наибольшее количество данных для обучения (85 слов, 2073 контекстов и 312 различных смыслов) разработать алгоритм, который позволит различать различные смыслы одного и того же слова по контексту (фактически, кластеризовать предложения на основе данных контекстов).

Далее, для того, чтобы оценить алгоритм различения смысла, предлагается запустить его на данных из дополнительной директории additional\active-rutenten\ , которая достаточно представительна с тз данных:

additional	Active Dict.	ruTenTen	train	21	7

в качестве критерия нужно использовать Adjusted Rand Index (ARI) и предлагаемый скрипт в репозитории.

Целью тестового задания является разработка алгоритма, работающего лучше базового подхода, adagram (код есть в репозитории, результаты также в каждой из папок представлены). Достичь в терминах ARI уровня не хуже 0.2. Код выполнения: python.

Как сказано в описании этого соревнования, возможны два подхода, в одном из которых используются тезарусы (sense inventories). В данном задании не предполагается подобного рода баз знаний. Однако, можно использовать словари, embeddings, и тп. вещи (включая технологии POS тегинга, синтаксической разметки и т.п) те речь идет о слеюущем треке:

- In the “knowledge-free” track participants need to induce a sense inventory from a text corpus of their own. The participants need to use it to assign each context with a sense identifier according to this induced inventory.