Методы решения задачи поиска похожих судебных решений

Ермолаев Павел Альбертович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: ermolaev.p.a@yandex.ru

Научный руководитель *— Полякова Ирина Николаевна*

С момента возникновения человеческого общества между людьми и группами людей неизменно происходили конфликты. В современном мире в любом правовом государстве межличностные и юридические отношения регулируются законодательством, и для разрешения конфликтов создается специальная государственная структура - судебная система, позволяющая в специальных учреждениях - судах - разрешать споры.

За продолжительную историю существования судов накопился большой объем информации по судебным тяжбам: заявления сторон, ход рассмотрения дела, решения суда, апелляции. Современные методы обработки больших объемов данных и компьютерная лингвистика позволяют извлечь, структурировать этот объем данных, а также применить полученную агрегированную информацию на благо людям.

Одной из задач, которую можно решить, обладая архивными данными по судебной практике, является задача поиска похожих судебных решений. Задача состоит в том, чтобы, имея текст решения суда или произвольный запрос по теме судебного решения, найти похожие судебные решения из некоторого множества уже имеющихся решений.

Поиск похожих судебных решений может использоваться при определении степени схожести судебных дел, при принятии решения о подаче апелляции и пр. Поиском пользуются юристы, адвокаты, судьи, простые граждане, не имеющим специального юридического образования. На первый взгляд может показаться, что задача поиска судебных решений сводится к задаче обычного органического поиска, однако это не так.

Проведенные эксперименты показывают, что из-за специфичности предметной области для эффективного поиска требуется особая предварительная обработка исходных документов и запроса, например, в данной задаче становится критичным корректное определение и удаление некоторых именованных сущностей (названий организа-

ций, ФИО, геолокаций). Также в этой задаче требуется использование нестандартных методов поиска, например примитивный полнотекстовый поиск основанный на tf-idf взвешивании word2vec векторов слов проигрывает поиску с выделением шаблонных частей в документах и последующему поиску в этих шаблонах. Так как структура судебных решений поддается шаблонизации, т.е. их структура в целом похожа и можно выделить общие важные фрагменты, например, обстоятельства дела, доказательства, вынесенное решение, то наиболее эффективным является выделение этих шаблонных фрагментов и их последующее независимое использование при поиске нужных фрагментов. Именно поиск по этим фрагментам, а не по всему документу дает наилучший результат.

Таким образом, поиск по судебным решениям плохо сводится к стандартной задаче поиска и требует более серьезной и осмысленной предварительной обработки, а также методов поиска, основанных в первую очередь на шаблонизации исходных данных, и поиска внутри найденных шаблонных структур.

Литература

1. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013): Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR.
2. Word2Vec, Doc2vec, GloVe: Neural Word Embeddings for Natural Language Processing
`https://deeplearning4j.org/word2vec.html`.
3. Pagliardini, M., Gupta P., Jaggi, M. (2018) Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. Proceedings of NAACL-HLT 2018, pages 528–540